2014-12-01

The Multidimensional Quality Metric (MQM) Framework: A New Framework for Translation Quality Assessment

Valerie Ruth Mariana
Brigham Young University - Provo

The Multidimensional Quality Metric (MQM) Framework:

A New Framework for Translation Quality Assessment

Valerie R. Mariana

A thesis 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
Anca Sprenger
Yvon LeBras

Department of French and Italian

Brigham Young University

December 2014

ABSTRACT

The Multidimensional Quality Metric (MQM) Framework:
A New Framework for Translation Quality Assessment

Valerie R. Mariana
Department of French and Italian, BYU
Master of Arts

This document is a supplement to the article entitled "The Multidimensional Quality Metric (MQM) Framework: A New Framework for Translation Quality Assessment", which has been acepted for publication in the upcoming January volume of JoSTrans, the Journal of Specialized Translation. The article is a coauthored project between Dr. Alan K. Melby, Dr. Troy Cox and myself. In this document you will find a preface describing the process of writing the article, an annotated bibliography of sources consulted in my research, a summary of what I learned, and a conclusion that considers the future avenues opened up by this research.

Our article examines a new method for assessing the quality of a translation known as the Multidimensional Quality Metric, MQM. In our experiment we set the MQM framework to mirror, as closely as possible, the American Translators Association's (ATA) translator certification exam. To do this we mapped the ATA error categories to corresponding MQM error categories. We acquired a set of 29 student translations and had a group of student raters use the MQM framework to rate these translations. We measured the practicality of the MQM framework by comparing the time required for ratings to the average time required to rate translations in the industry. In addition, we had 2 ATA certified translators rate the anchor translation (a translation that was scored by every rater in order to have a point of comparison). The certified translators' ratings were used to verify that the scores given by the student raters were valid. Reliability was also measured, which found that the student raters were not interchangeable, but that the measurement estimate of reliability was adequate.

The article's goal was to determine the extent to which the Multidimensional Quality Metric framework for translation evaluation is viable (practical, reliable and valid) when designed to mirror the ATA certification exam. Overall, the results of the experiment showed that MQM could be a viable way to rate translation quality when operationalized based on the ATA's translator certification exam. This is an important discovery in the field of translation quality, because it shows that MQM could be a viable tool for future researchers. Our experiment suggests that researchers ought to take advantage of the MQM framework because, not only is it free, but any studies completed using the MQM framework would have a common base, making these studies more easily comparable.

Keywords: Translation quality, translation evaluation, translation assessment, Multidimensional Quality Metric, MQM, practicality, validity, reliability

ACKNOWLEDGEMENTS

I would like to thank Dr. Melby for spearheading this project, despite his impending retirement, and Dr. Cox for adding his statistics prowess to the paper. My gratitude also goes out to Yvon LeBras for being on my committee, and especially to Anca Spenger, who was infinitely helpful, both as a committee member and in her role as the graduate coordinator. My husband, Devin, deserves the greatest thanks for supporting me through the long hours required by this program.

Contents

Introduction

In accordance with thesis option two, Dr. Alan K. Melby and I co-authored the article "The Multidimensional Quality Metric (MQM) Framework: A New Framework for Translation Quality Assessment" with Dr. Troy Cox joining us to increase the credibility of the article via his extensive knowledge of the statistics necessary for determining reliability. The article has been accepted by JoSTrans for publication in the upcoming volume of the journal, which is set for release in January of 2015.

My goal in entering the French MA program was to become better acquainted with the French language and culture so as to become a better translator. Within the first few weeks of joining the program, I contacted Dr. Melby, with whom I had had a French translation course the year before, and asked him if he had any projects with which I could be involved. At that point we began meeting together regularly, discussing what later became our experiment on MQM and the article that resulted. For many weeks we met in person to define the ever-metamorphosing goal of the experiment, while I learned more about what exactly MQM was and why it was important in the world of translation. Dr. Cox joined us in meetings consisting of various professors and students who met to discuss translation quality and MQM.

As I did the research necessary for this paper, I found that it was vital to take notes on everything I read, even if I did not think it to be useful at first glance. The framework provided by Anca Sprenger for note-taking was vital. Things that I had read over a year ago were easy to access and recall due to my detailed notes. It was especially important to note how to find the source again, which prevented hours of frantic searching for a lost article to cite. Dr. Melby and Dr. Cox were both very helpful in pointing me in the right direction to find useful ideas, key

definitions and thought provoking articles. In addition, Dr. Melby and Dr. Cox both had very different insights and were able to help me with different aspects of the research. Clearly, a well-rounded paper takes many different people who have different specialties. Each person's specialty allows him or her to contribute to the base research of paper in a unique and invaluable way. For this reason, much more collaboration was required than I would have imagined at first. Communication was, therefore, of upmost importance.

I began writing drafts of the planned experiment, which I would email to Dr. Melby and which he would revise using the "Track Changes" tool in Microsoft Word. Overall, I wrote at least 75% of the text, and was the one to implement any suggested changes. Later on in the semester, Dr. Melby revealed his upcoming retirement and we sped up the process as much as we could, though at the same time, Dr. Melby was often traveling to conferences. During this time period, I took an introduction to psychometrics class from Dr. Cox that focused on the Rasch measurement which was used in the reliability section of the article. As I communicated with Dr. Melby via email, I was able to see Dr. Cox in person. With the help of Dr. Melby I created the means by which my raters could use MQM to rate the student translations. With the help of Dr. Cox, I created a rating schedule and training materials.

Over the summer, I met with the raters for a pilot of the rating materials, following which I trained all of the raters. Some raters were in different states or overseas, so training was conducted in person, over the phone, via videoconference or via email, depending on the geographical distance between the rater and myself. Eventually, ratings were completed by all of the raters, and I transformed the data from percentages to scores from 0 to 9. These scores were then analyzed by the program Facets (Linacre, 2013) to determine the reliability. The results of

the Facets analysis showed the raters to not be interchangeable. In other words, they displayed low inter-rater reliability.

Dr. Cox had mentioned to me that this was not a problem, and that the system could still be considered reliable. However, Dr. Melby and I did not fully understand why this was not a problem and wanted to get Dr. Cox together with Dr. Fields to come to a consensus on reliability. However, we were unable to consult further with Dr. Cox about this, as he was out of town. Dr. Melby took the results to Dr. Fields, who manipulated the data further. We put Field's information into the submitted paper.

However, after getting back in touch with Dr. Cox, and doing more research on reliability, I have found that there are multiple types of reliability, and though our consensus estimate of reliability was low, our measurement estimate of reliability was acceptable. In a measurement estimate of reliability is possible to calculate a fair average score for the student translators. This fair average adjusts for the differences between the raters and calculates the score the student ought to have gotten if the raters were interchangeable. Had I realized the differences between these types of reliability before, we would not have needed to reinterpret the data and could have easily published what we initially found. Between submitting the paper and having it be accepted months later, we were able to correct the reliability section.

Practicality was determined via a comparison of the time it took to rate these translations and the average time required for the translation industry.

In order to examine validity, Dr. Melby contacted two ATA certified translators who agreed to rate the anchor translation, and whose scores we compared to the student raters' scores. This method for determining validity was suggested by Dr. Cox and is known as the interpretive argument method, as described by Kane (Kane, 2006). This method takes a look at established

facts and makes an argument for validity based on those facts. In a nutshell, our argument was that since our novice raters were statistically similar to the professionals, their work was valid.

Annotated Bibliography

"A Comparison of Norm-referencing and Criterion-referencing Methods for Determining Student Grades in Higher Education. "*Assessing Student Learning*. Centre for the Study of Higher Education, 2002. Web. 19 July 2014.

This article offers another description of the difference between norm-referenced and criterion-referenced, this time from the Australian perspective. It judges norm-referencing to be unfair on its own and criterion-referenced tests to be much preferable. Ideally, higher education ought to find a balance between the two, leaning more towards criterion-referencing.

"ATA Certification Program Rubric for Grading." *Atanet.org*. American Translators Association, 2011. Web. 21 Jan. 2014.

The ATA uses this rubric in its translator certification exams. The rubric gives useful information on the error categories that the ATA recognizes, which was helpful in designing our rating scorecard. This is an example of an analytic approach to rating translations in a translation education environment.

Bachman, Lyle F., and Adrian S. Palmer. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford UP, 1996. Print.

This document defines usefulness, stating that a test is useful if it has good reliability, construct validity (especially useful for testing), authenticity, interactiveness, impact, and practicality. Every test designer should maximize overall test usefulness, not individual qualities. As such, qualities are not evaluated independently, instead an evaluator should look at their combined effect on usefulness. The balance between qualities depends on

the specific situation. Reliability is defined as consistency of measurement. For example, if the student gets a score of 90% on version 1, that student should also get a 90% on version 2. Validity asks if the test score means something significant. A test must measure what the designers want it to and nothing else. Authenticity measures whether good performance on a test actually means the test-taker is good at that skill in real life. Interactiveness is how the test interacts with language ability, topical knowledge and affective schemata (the degree to which the constructs we want to assess are involved in accomplishing the test). Impact is simply the impact the test scores have on the test taker and on society. Practicality is whether implementing the test is feasible (practical) in terms of resources, for example money, personnel and time. Overall this article gave great definitions of the qualities we wanted to consider in this paper, although interactiveness and impact were deemed unimportant for our study. Our omission of them is, however, still supported by this article in that it allows us to maximize usefulness.

Colina, Sonia. "Further Evidence for a Functionalist Approach to Translation Quality Evaluation." *Target: International Journal on Translation Studies* 21.2 (2009): 235-64. *EBSCOhost*. Web. 16 Sept. 2013.

Colina created a translation evaluation tool designed for use in hospitals that would be "componential, functionalist, textual… [which] evaluates components of quality separately … [and is] relative to the function and the characteristics of the audience specified for the translated text" (240). This paper focuses on the second round of testing for the tool to see if it has "good inter-rater reliability" (241). Colina's "proposed TQA tool allows for a user-defined notion of quality in which it is the user or requester who decides which aspects of quality are more important " (240). This is similar to our idea of

specifications, wherein the client decides what is most important to the text, which in turn determines quality.

Conde, Tomas. "Translation Evaluation on the Surface of Texts: A Preliminary Analysis." *The Journal of Specialized Translation* 15 (2011): 69-86.*JoSTrans: The Journal of Specialized Translation*. JoSTrans, Jan. 2011. Web. 19 July 2014.

In this article, Conde states that when evaluating a translation there are independent factors that affect the evaluation, such as whether the errors are in an emphasized part of the text (located in headings etc.) or whether the formatting is off. Such errors might be penalized more severely, due to their being more obvious. In addition, a translation that is shorter than the others might be unconsciously penalized. Furthermore, he sates that the two most important independent variables are consistency and productivity (finishing the translation). Conde goes on to give a very nice definition of the analytic method, saying that it is a system where errors are counted and characterized. The most widely used criteria of this characterization of errors is the nature and importance or the error. This matches with the MQM scorecard, which also focuses on the nature (error category) and importance (level of severity) of the errors. Conde remarks that analytical evaluation systems are still the most widely used, especially within teaching environments.

Eckes, Thomas. *Introduction to Many-facet Rasch Measurement: Analyzing and Evaluating Rater-mediated Assessments*. Frankfurt Am Main: Peter Lang, 2011. Print.

This book describes the basics of Rasch measurement, defining the differences between dichotomous and polytomous, discussing the reasons for using many facets and describing various rating designs. Rasch measurement can determine rater variability and inter-rater reliability. There are several types of inter-rater reliability which include

consensus where all raters give essay 1, for example, a score of 7, consistency where all raters rank essay 1 as the 2nd best, though they may not give it the same score. How to calculate inter-rater reliability is described. Interestingly the book states that "…rater training usually does not succeed in reducing between-rater severity differences to an acceptably low level. Therefore, in most situations, adopting the standard view that rater training needs to pursue to objective of achieving maximal between-rater similarity and acting accordingly in rater training sessions, is highly likely to fail. The constructive alternative ... is to accept rater heterogeneity as a fact of life..." p55. It does however go on to say that rater-training can make raters more internally consistent which makes calculating a fair score with the Rasch model more effective.

Evans, Norman W., K. James Hartshorn, Troy L. Cox, and Teresa Martin De Jel. "Measuring Written Linguistic Accuracy with Weighted Clause Ratios: A Question of Validity." *Journal of Second Language Writing* 24 (2014): 33-50. *Www.sciencedirect.com*. Science Direct. Web. 16 Apr. 2014.

This article presents an interesting definition of error. Basically, an error is something a native speaker would not say, write or do. The article gives a summary of a bunch of ways to record errors. One will notice that the PIE ratings in our study are similar to the error-free T units. Each T unit is basically a single sentence, so the segments used in our MQM rating are T units. The article goes on to describe first, second and third-degree errors, which correspond to minor, major and critical errors respectively. The article contains definitions of construct validity, criterion-related validity, and content validity. It includes a good explanation of the agreement-accuracy paradox where though rater's may agree on ranking students from worst to best, they might use the scale differently, with

one always awarding a slightly higher score, or they may agree perfectly, but really just both be using the scale incorrectly. Thus we see that agreement and accuracy do not always follow one another. Taking this paradox into account, one way to account for the systematic variation in scores is to use Many-Facet Rasch Measurement (MFRM). The article describes what MFRM is and how it can be used. It includes information on how to convert numbers to run a MFRM on them and how to interpret the data.

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

This article discusses a translation evaluation method called CDI (calibration of dichotomous items), which is similar to PIE. This article compares holistic, analytic and CDI methods. It gives a brief history of translation testing, and speaks on how people are starting to want more reliable and valid testing methods. CDI was developed as a norm-referenced test “with the aim of freeing translation assessment from construct-irrelevant variables that arise in both analytic (i.e. evaluating by means of pre-conceived criteria) and holistic…scoring" p 75. The authors used holistic, analytic and CDI methods to grade 100+ translations. CDI was found to be the most reliable, while holistic and analytic methods were found to be more subjective and less reliable.

Hallgren, Kevin A. "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial." *Tutor Quant Methods Psychol.* 8.1 (2012): 23-34. *National Center for Biotechnology Information*. U.S. National Library of Medicine, 18 Dec. 2012. Web. 01 Sept. 2014.

Inter-Rater Reliability quantifies the "degree of agreement" between raters. Classical test theory states that a person's observed score is equal to their true score adjusted for measurement error and one source of such error can be poor inter-rater reliability. Inter-rater reliability analysis aims to determine how much of the variance in the observed scores is due to variance in the true scores after the variance due to measurement error between coders has been removed. The article notes that in a rating design that is not fully crossed (in other words, not every rater grades every translation) it may be necessary to use additional statistics to prove reliability. It is also important to make sure the scale doesn't have a restricted range (ie. All scores are 4s and 5s rather than an even spread between 0 and 5) as this may lower reliability. Carefully training raters is also imperative to increasing inter-rater reliability. The article advises against reporting the percentage of times to raters agree as proof of inter-rater reliability as some agreement would be by random chance, so this method overestimates agreement. In addition, a researcher must remember to report the statistic used to determine inter-rater reliability, which must be carefully selected to be right for the study. In addition, a good researcher will interpret the inter-rater reliability statistic, expanding on its implications, and explaining why inter-rater reliability might be low.

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

This article gives the definition of validity that we adopted for our paper. It argues that validity is based on an argument that has propositions that support the interpretation of the argument and states the evidence that supports the propositions. There are many types of evidence and validity depends on how this evidence supports the interpretation of test

scores. For example, a proposition might be that content in the test is relevant to the industry, which might be supported by the evident of empirical studies. The best evidence is extensive and draws largely upon current literature.

Koby, Geoffrey S., and Alan K. Melby. "Certification and Job Task Analysis (JTA): Establishing Validity of Translator Certification Examinations." *The International Journal of Translation and Interpreting Research* 5.1 (2013): 174-210. *Translation & Interpreting*. TransInt, 2013. Web. 19 July 2014.

Most certifying organizations include an examination. ISO 1700 standards regulate certification. Next there is a discussion of how a test could be reliable but not valid and vice versa. In this article, an examination is valid when you are testing what you want to test. ATA used a Job Task Analysis to prove validity. An examination is reliable when the candidate gets the same score, within a reasonable range of variation, regardless of who grades the examination. The author discusses the necessity of a "translation brief", which is essentially the same as translation specifications. The article defines competence as a combination of knowledge, skills and abilities, also know as KSAs. A list of various KSAs are presented, then ranked in importance. The ATA competency categories matched what were in the literature. The ATA's KSAs were sorted by professionals into 3 categories (1. Prerequisites to take exam, 2. Exam 3. Professional development). These KSAs must be considered when developing a translator certification examination, in identifying the necessary knowledge, skills and abilities needed for a translator, then creating an examination that actually measures these things. This makes the test valid. However, it is still necessary to demonstrate the reliability of the grading.

Koby, Geoffrey, and Gertrud Champe. "Welcome to the Real World: Professional-Level Translator Certification." *The International Journal of Translation and Interpreting Research* 5.1 (2013): n. pag. Print.

This article details how the ATA translator certification test works. It discusses the different error categories, which is useful information for our paper.

“Le Prof "désobéisseur", Sanctionné, Perd 7.000 Euros." *Nouvelobs.com*. Le Nouvel Observateur, 24 July 2009. Web. 14 Nov. 2014.

A modified version of this article was the French source text for the translations our raters rated in this study.

Linacre, John M. "Misfit Diagnosis: Infit Outfit Mean-square Standardized." *Www.winsteps.com*. Winsteps Rasch Measurement Software, n.d. Web. 16 Nov. 2014.

This article shows how to interpret the statistics output by the Facets program, specifically outfit mean square. Outfit mean square measures the self-consistency of raters, which was important to the reliability aspect of our study. All of our raters were within the acceptable range for outfit mean square, as defined by this article.

"Multidimensional Quality Metrics (MQM) Definition." *QTLaunchPad*. Ed. Arle Lommel. Quality Translation Launch Pad, 2014. Web. 06 Sept. 2014.

This document categorizes the issue types that may be found in a translation. Issues are something that might be wrong and they turn into errors if they are incorrect, but may be okay if done for an explicit purpose by the translator. Errors can be weighted or simply counted to give the translation a score. The categories of accuracy, fluency, verity, design, internationalization, compatibility and all of their extensions are defined. The

definitions presented on this site were useful in understanding the MQM version of assessing translation quality and in building the scorecard for the study.

Nord, Christiane. *Translating as a Purposeful Activity: Functionalist Approaches Explained.* Manchester, UK: St. Jerome Pub., 1997. Print.

Nord's book talks about the functionalist approach to translation, which is a precursor to the idea of translation specifications, a predetermined set of requirements that the translator and client agree on so that the translation fits the client's needs. Translation specifications are important in defining translation quality and functionalist translation is a necessary part of the discourse on translation quality, which must be touched upon.

O'Brien, Sharon. "Towards a Dynamic Quality Evaluation Model for Translation." *The Journal of Specialized Translation* 17 (2012): 55-77. *JoSTrans: The Journal of Specialized Translation*. JoSTrans, Jan. 2012. Web. 19 July 2014.

In this article TAUS, the Translation Automation Users Society is looking for a new translation evaluation method that fits budgetary constraints, new paradigms, new technology and new focus. New paradigms refers to the nature of text itself is changing with the invention of tweets, multimedia and user-generated content. New technology entails a better integration of MT or TMs and new focus refers to how companies focus more on the end users' perceptions of quality than before. TAUS did a benchmarking exercise on 11 existing Quality Evaluation methods. They found similarities and differences between the methods. Generally in the industry 3-4 minor errors per thousand words or 1 major error per thousand words was found to be acceptable. Minor errors were defined as errors that do not confuse the reader, whereas major errors take a large toll on understanding, and critical errors may affect the usefulness of the translation, or cause

safety or liability concerns. Certain languages, such as Japanese and French had a lower error tolerance. Overall the industry prefers to use error-counting methods for evaluation. This article was a good source for defining minor, major and critical errors as well as a good source for possible error categories as well as a great description of analytic evaluation methods.

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

This book presents simple definitions of validity and reliability. It examines the differences between equivalence reliability, test-retest reliability, and inter-rater reliability. It also discusses the differences between the types of validity: face validity, content validity, concurrent validity, and predictive validity. In addition, the book reveals the difference between norm referenced testing and criterion referenced testing. Norm referenced testing orders people from best to worst using the test. A person's score is based on how well everyone else does. Criterion referenced tests assign scores based on what the examinee can do, regardless of how well everyone else does. On a criterion-referenced test, everyone can get an A. The book goes on to describe different types of tests and methods of test construction.

Snow, Tyler. "Establishing the Viability of the Multidimensional Quality Metrics Framework." Thesis. Brigham Young University, Forthcoming. Print.

Snow's thesis gave us the information we needed to determine practicality for this study. He sent out a survey to a number of players in the translation industry and found out the average time it took in the industry for translations to be rated.

Stemler, Steven E. "A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability." *Practical Assessment, Research & Evaluation* 9.4 (2004): n. pag. *Practical Assessment, Research & Evaluation: A Peer Reviewed Electronic Journal*. PARE, 2004. Web. 01 Sept. 2014.

This article recognizes 3 branches of inter-rater reliability. The first is the consensus estimate which is "based on the assumption that reasonable observers should be able to come to exact agreement about how to apply the various levels of a scoring rubric…" and is used for nominal data. There are various means for calculating the consensus agreement, one of the most common being Cohen's kappa. There are disadvantages for this method, including the fact that each reliability must be calculated for each pair of judges on each item, and the amount of time it takes to train judges to be exactly the same. Our study proved that the raters did not display adequate reliability in the consistency estimate. The second type is consistency estimates where "each judge is consistent in classifying the phenomenon according to his or her own definition of the scale." This estimate works well for contiguous data. It can be calculated with the Pearson correlation coefficient, the Spearman rank coefficient or Cronback's alpha. It did not apply to this study. The last type of inter-rater reliability is the measurement estimate for which "it is not necessary for two judges to come to a consensus on how to apply a scoring rubric because differences in judge severity can be estimated and accounted for in the creation of each participant's final score. " This method is used when multiple judges can't all rate everything, which was the case with our study. The measurement estimate of inter-rater reliability can be found via various techniques, including the many-facets Rasch model, which is what we used in our study. Rasch "allows judge severity to be

derived using the same scale (i.e., the logit scale) as person ability and item difficulty." The advantages of the measurement estimate include the fact that the summary scores are more accurate than the regular scores, it handles multiple judges and multiple items at once, and not all judges have to rate all items.

"The Length of a Logit." On line at:http://www.rasch.org/rmt/rmt32b.htm

This article explains the scale used in the output tables of the Facets program. Simply put, a logit is a unit of measurement that allows us to compare many diverse, otherwise incompatible Facets to one another on the same scale.

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

Vermeer is the founder of the discussion on translation quality, with his skopos theory, a precursor for our translation specifications. Considering the importance we place on specifications, we found it necessary to at least mention skopos theory and its similar way of defining translation quality.

Waddington, Christopher. "Different Methods of Evaluating Student Translations: The Question of Validity." *Meta: Translators' Journal* 46.2 (2001): 311-25. *Érudit*. Érudit, 2001. Web. 16 July 2013.

In this article, the author examines the validity the of results of 4 different evaluation methods that fall under that categories of analytic, rubric and holistic, which have been used to evaluation the quality of student translations. Each method's results were compared to 17 external criteria (including intelligence tests, self-assessments, teacher assessments, grades in the class, grades on other exams). The author defines translation

competence as the ability to understand and transfer source text in addition to the ability to express it in the target language. The results of this study showed that each method was valid.

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

This article details the advantages to using many-faceted Rasch measurement, which is what the program Facets uses. It explains how the statistics can be used to correct for judge bias if necessary, in order to find the score the examinee ought to have gotten. This information was useful for our discussion of reliability.

Conclusion

This study opened up many avenues for future research. Including an in-depth study of the Pre-Selected Items Evaluation method, PIE, operationalized in the same manner as was MQM in this study. We originally intended to include PIE, which we considered to be a subcategory of MQM, in this study, to see if it predicts the full score given by MQM. In addition, this study lays the groundwork for another study on MQM using a more appropriate source text and focusing more heavily on rater training. It would also be useful to test the method on different language pairs. What's more, we hope this article encourages other researchers to use the MQM framework in their own studies, provided that they take caution in training the raters and take into account the need to run results through a program like Facets that can correct for judge differences, as is necessary for measurement estimates of reliability. MQM could be a powerful uniting force for the field of translation-related research, because it can be personalized to fit different projects, and studies using it should be fairly comparable since they will have the common base of the MQM framework.

Bibliography

"A Comparison of Norm-referencing and Criterion-referencing Methods for Determining Student Grades in Higher Education."*Assessing Student Learning*. Centre for the Study of Higher Education, 2002. Web. 19 July 2014.

“ATA Certification Program Rubric for Grading.” *Atanet.org*. American Translators Association, 2011. Web. 21 Jan. 2014.

Bachman, Lyle F., and Adrian S. Palmer. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford UP, 1996. Print.

Colina, Sonia. "Further Evidence for a Functionalist Approach to Translation Quality Evaluation." *Target: International Journal on Translation Studies* 21.2 (2009): 235-64. *EBSCOhost*. Web. 16 Sept. 2013.

Conde, Tomas. "Translation Evaluation on the Surface of Texts: A Preliminary Analysis." *The Journal of Specialized Translation* 15 (2011): 69-86.*JoSTrans: The Journal of Specialized Translation*. JoSTrans, Jan. 2011. Web. 19 July 2014.

Eckes, Thomas. *Introduction to Many-facet Rasch Measurement: Analyzing and Evaluating Rater-mediated Assessments*. Frankfurt Am Main: Peter Lang, 2011. Print.

Evans, Norman W., K. James Hartshorn, Troy L. Cox, and Teresa Martin De Jel. "Measuring Written Linguistic Accuracy with Weighted Clause Ratios: A Question of Validity." *Journal of Second Language Writing* 24 (2014): 33-50. *Www.sciencedirect.com*. Science Direct. Web. 16 Apr. 2014.

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

Hallgren, Kevin A. "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial." *Tutor Quant Methods Psychol.* 8.1 (2012): 23-34. *National Center for Biotechnology Information*. U.S. National Library of Medicine, 18 Dec. 2012. Web. 01 Sept. 2014.

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

Koby, Geoffrey S., and Alan K. Melby. "Certification and Job Task Analysis (JTA): Establishing Validity of Translator Certification Examinations." *The International Journal of Translation and Interpreting Research* 5.1 (2013): 174-210. *Translation & Interpreting*. TransInt, 2013. Web. 19 July 2014.

Koby, Geoffrey, and Gertrud Champe. "Welcome to the Real World: Professional-Level Translator Certification." *The International Journal of Translation and Interpreting Research* 5.1 (2013): n. pag. Print.

“Le Prof "désobéisseur", Sanctionné, Perd 7.000 Euros." *Nouvelobs.com*. Le Nouvel Observateur, 24 July 2009. Web. 14 Nov. 2014.

Linacre, J. M. (2013) Facets computer program for many-facet Rasch measurement, version 3.71.2. On line at: www.winsteps.com

Linacre, John M. "Misfit Diagnosis: Infit Outfit Mean-square Standardized." *Www.winsteps.com*. Winsteps Rasch Measurement Software, n.d. Web. 16 Nov. 2014.

"Multidimensional Quality Metrics (MQM) Definition." *QTLaunchPad*. Ed. Arle Lommel. Quality Translation Launch Pad, 2014. Web. 06 Sept. 2014.

Nord, Christiane. *Translating as a Purposeful Activity: Functionalist Approaches Explained.* Manchester, UK: St. Jerome Pub., 1997. Print.

O'Brien, Sharon. "Towards a Dynamic Quality Evaluation Model for Translation." *The Journal of Specialized Translation* 17 (2012): 55-77. *JoSTrans: The Journal of Specialized Translation*. JoSTrans, Jan. 2012. Web. 19 July 2014.

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

Snow, Tyler. "Establishing the Viability of the Multidimensional Quality Metrics Framework." Thesis. Brigham Young University, Forthcoming. Print.

Stemler, Steven E. "A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability." *Practical Assessment, Research & Evaluation* 9.4 (2004): n. pag. *Practical Assessment, Research & Evaluation: A Peer Reviewed Electronic Journal*. PARE, 2004. Web. 01 Sept. 2014.

“The Length of a Logit.” On line at:http://www.rasch.org/rmt/rmt32b.htm

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

Waddington, Christopher. "Different Methods of Evaluating Student Translations: The Question of Validity." *Meta: Translators' Journal* 46.2 (2001): 311-25. *Érudit*. Érudit, 2001. Web. 16 July 2013.

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