The Language Product Evaluation Tool: Establishing Standards and Developing Workforce Expertise

Erica B. Michael[1], Allison Blodgett[1], Dominic Massaro[1,2], Benjamin Bailey[1,3], Diane de Terra[1], Scribner Messenger[4], Lelyn Saner[1], Kathy Rhoad[1], Shaina Castle[1], & Solveig Gannon-Kurowski[1,3]

[1]University of Maryland Center for Advanced Study of Language, USA

[2]University of California, Santa Cruz, USA

[3]University of Massachusetts, Amherst, USA

[4]United States Department of Defense

Corresponding Author

Erica B. Michael, PhD

University of Maryland

Center for Advanced Study of Language

7005 52nd Ave

College Park MD 20912 USA

emichael@casl.umd.edu

(+001) 301-226-8868

Biography

Erica B. Michael is an Associate Research Scientist at the University of Maryland Center for Advanced Study of Language (CASL). She is also affiliated with the University's Second Language Acquisition Program and Department of Psychology. Her training is in cognitive psychology and psycholinguistics, and she received her PhD in psychology in 1998 from Penn State University. Prior to joining CASL's research staff in 2005, Dr. Michael received postdoctoral training at Carnegie Mellon University and served as a visiting assistant professor at Bryn Mawr College. Her research interests include lexical and semantic processing in bilinguals and second language learners, and her CASL work focuses on cognitive processing in language and analysis tasks such as translation and summarization.

0. Abstract

The Language Product Evaluation Tool (LPET) offers a framework for evaluating a range of language products, including full translations and various types of summaries. The result of a collaborative effort between researchers and practitioners working in a US government setting, the LPET serves as a practical tool aimed at establishing translation standards; facilitating the quality review process and the delivery of consistent, systematic feedback; providing information for planning professional development activities for translators; and providing aggregated data describing individual and organizational capability and performance. We begin this paper by describing the development of the LPET and its major features. We then highlight findings from operational testing with quality reviewers and describe validity testing that is in progress. We end by presenting a sample view of aggregated LPET data of the type that might be useful at the managerial or organizational level.

1. Introduction to the Language Product Evaluation Tool (LPET)

1.1. Motivation for the LPET. A critical task for language analysts in the United States (US) government is translation of foreign language materials (both written and spoken) into written English. Language analysts produce a range of language products, including full translations, complete summaries, targeted summaries, gists, and hybrids that combine features of multiple product types. Although it is common for products to be reviewed by a second, typically more experienced, language analyst, there is no government-wide standard for the precise nature of the review. As a result, practices for reviewing and providing feedback vary across and sometimes within organizations (Michael et al., 2008). Standardized assessments are vital to ensuring high standards of quality review, giving language analysts meaningful feedback for improvement, and providing managers with a set of metrics to help them understand and improve the capability and performance of their workforce.

1.2. Precursor to the LPET. In response to this need to establish uniform translation standards within the US government context, a team of researchers at the University of Maryland Center for Advanced Study of Language (CASL) first developed an analytic rubric aimed at explicating the components required for one particular type of language product referred to in government contexts as *targeted summary translation*. Targeted summaries are typically written

in response to customer requirements (akin to “translation briefs” or “commissions” in non-government contexts). For example, a targeted summary of a speech by a Chinese Communist Party official would be very different if it were written in response to a request for information about plans to float the Chinese *yuan* than if it were written in response to a request for information about Chinese relations with North Korea.

In developing an analytic rubric for targeted summaries, we consulted the scientific literature; conducted interviews with language analysts, area specialists, quality reviewers, and instructors; and evaluated the strengths and weaknesses of an earlier holistic rubric (Michael, Bailey, Gannon-Kurowski, & Pinckney, 2007). Following standards in educational and psychological measurement (e.g., Crocker & Algina, 2006), we strove to produce a rubric characterized by the following qualities:

- All elements worthy of evaluation are included.
- Each element is unidimensional in that it does not overlap with other elements.
- Each element communicates clearly to the user.
- The rating on each element covers the range of performance, typically in the range of 3 to 7 levels.

The resulting analytic rubric, known as the Summary Translation Evaluation Tool (STET), encapsulates six analytical dimensions that cover a summary’s content, structure, and style: Significance, Completeness, Accuracy, Omission of Irrelevant Information, Organization, and Writing. (For a more detailed description of the STET and its six dimensions, see Michael, Massaro, & Perlman, 2009.)

To conduct a rigorous evaluation of the STET, we created a set of summaries established by experts to represent specified levels of performance along the six dimensions; a group of 38 language analysts then used the STET to evaluate each summary (Michael et al., 2010). Quantitative modeling of the results suggested that the tool was valid and sensitive; that is, variations in summary quality were appropriately reflected in ratings for the corresponding dimension and the STET permitted a reasonable range of performance ratings along each dimension. Reliability testing demonstrated some consistency across users and conditions but suggested a need for more extensive training, a finding that was instrumental in establishing the training protocol for the LPET.

1.3. Development of the LPET. Targeted summary translations are increasingly common in government contexts because of the overwhelming amount of material to be processed, making it impractical and unnecessary to write full translations of all source items. However, full translations and other types of language products are still required in many situations, and the LPET was developed because government practitioners who encountered the STET asked for a similar tool that extended to all product types. Like the STET, the LPET is a practical tool aimed at (1) establishing translation standards; (2) facilitating the quality review process and the delivery of consistent, systematic feedback; (3) providing information for planning professional development activities for translators; and (4) providing aggregated data describing individual and organizational capability and performance.

Although academe and industry have already developed a variety of models for translation assessment, many of them fail to achieve the combination of practical applicability and theoretical soundness that is embodied in the STET and the LPET. Colina (2008, 2009) succinctly captured the issues on both sides: On the one hand, practical translation evaluation schemes, such as the Society of Automotive Engineers translation quality metric for the automobile industry (SAEJ2450) or the American Translators Association certification test grading criteria,[1] have been developed *ad hoc* based on common patterns of errors, so they consist largely of lists of such errors that can be checked off. Because these schemes are based on neither an explicit theory of translation nor conceptually-based categories, they do not provide orienting principles for translation assessment that might be adapted for use in other contexts. On the other hand, many theoretical models of translation have been developed without reference to practical applicability. Such theories pose foundational questions about the type or degree of equivalence between source and target texts and the relative privileging of author, translator, or audience, but they fail to specify on-the-ground categories or metrics for deciding whether a given translation satisfies the abstract criteria of a given theory. The basic features of the LPET, presented in the next section, capture conceptual categories that underpin language product quality while providing a concise and systematic way for reviewers to document product quality and task difficulty.

1 See http://standards.sae.org/j2450_200508/ and https://www.atanet.org/certification/aboutexams_error.php, respectively.

1.4. Basic Features of the LPET. Figure 1 provides a snapshot of the LPET, the development of which reflects two guiding principles: First, the quality of a language product is a composite of the three main components seen on the right-hand side of the form: *Language Performance*, *Analysis*, and *Presentation*. In other words, quality is a function of more than just how accurately the source language is rendered into English. In many cases a language product must reflect an additional level of analysis (beyond the analysis inherent in translation) to ensure that the text conveys the significance of the source material and provides additional contextual information. The use of accepted conventions for format and writing is also important for a language product to communicate clearly to its reader.

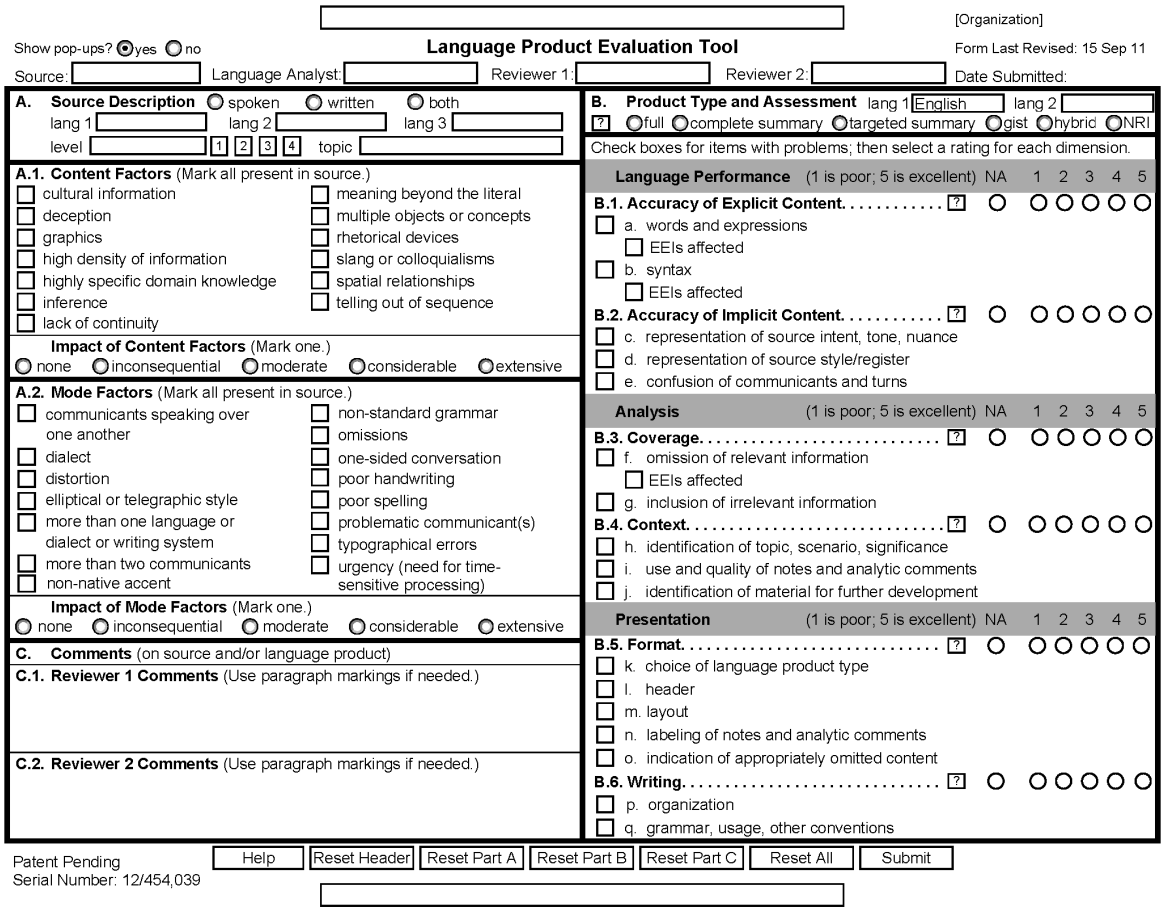

Figure 1. The right-hand side of the LPET focuses on product type and assessment of quality. It highlights three components (*Language Performance*, *Analysis*, and *Presentation*) and provides rating scales for six essential dimensions (*Accuracy of Explicit Content*, *Accuracy of Implicit Content*, *Coverage*, *Context*, *Format*, and *Writing*). The left-hand side of the LPET focuses on characteristics of the source material that may affect difficulty level, and provides space for reviewer comments. The electronic version of the LPET provides the user with pop-up windows containing descriptions and examples of the individual elements.

The three main components each comprise two essential dimensions, with a rating scale for each of those dimensions. The quality of *Language Performance* is a function of *Accuracy of Explicit Content* and *Accuracy of Implicit Content*; the quality of *Analysis* is a function of *Coverage* and *Context*; and the quality of *Presentation* is a function of *Format* and *Writing*. The relative importance of each dimension is not pre-specified and will depend on the purpose of the language product.

Each dimension has a corresponding set of checkbox items, allowing reviewers to identify particular problem areas that contribute to each rating. These checkbox items represent common error types and features that are especially important to product quality within the US government context. For example, the dimension *Accuracy of Explicit Content* includes checkboxes for indicating problems in the rendering of *words and expressions* and *syntax,* and in each case allows the reviewer to indicate whether or not the problems affect the rendering of the essential elements of information.

The second guiding principle behind the LPET is that the quality of a language product should be understood in the context of the difficulty of the source material, captured on the LPET's left-hand side. For example, a language product that is missing key information may reflect distortions in the source rather than the skill level of the language analyst who created the product. The LPET prompts users to document a variety of source characteristics, including source language(s), passage level,[2] and topic. Users further select checkboxes to indicate the presence of common features that can contribute to task difficulty, including both content factors, such as *cultural information* or *high density of information*, and mode factors, such as *dialect* or *poor spelling*.

1.5. LPET Training. To ensure appropriate and reliable use of the LPET, all quality reviewers who intend to use the LPET on the job are strongly encouraged to attend a 1-day workshop. The core of the LPET workshop is a series of exercises in which attendees work through the various components of the LPET as they evaluate sample language products that demonstrate particular features. Each sample language product is part of a package containing the source material, a model language product, a request for information (if the product is a targeted summary), and one or more additional products in which tracked changes point to areas

[2] See www.nflc.org/projects/recent_projects/passage_rating or www.govtilr.org.

needing improvement. Attendees learn to identify the appropriate checkbox for each problem within a language product and to assign a rating to each of the LPET product assessment scales.

1.6. Flexibility of LPET Use. The LPET was designed to allow flexible use with diverse source material and product types. For example, some of the mode factors describing source material are specific to written material (e.g., *poor handwriting*), some are specific to spoken material (e.g., *non-native accent*), and some apply to both (e.g., *non-standard grammar*). Similarly, not all of the product assessment items are appropriate for all product types. For example, *organization* applies to various types of summaries but not to full translations because in a full translation the language analyst simply maintains the organization of the source material. The LPET is also flexible in that it can be used in a variety of contexts, including both classroom and operational settings. When used to conduct quality review on the job, not every language product requires a completed LPET. Rather, the LPET is intended to be used flexibly, and reviewers may adjust the frequency of use in accordance with any number of factors, such as product type, product length, and experience levels of the analyst and reviewer.

2. Operational Testing of the LPET

2.1 Goals. Small groups of quality reviewers pilot tested the LPET to ensure that it meets the needs of practitioners. The pilot tests were designed to gather input on how the LPET works in authentic operational environments, to develop recommendations for improving the LPET and its training, and to begin exploring ways in which data might be aggregated for use by reviewers, analysts, managers, and instructors.

2.2. Methods. Four reviewers at each of two US government organizations participated in LPET training and then used the LPET for five to six months to review selected language products. Data included the completed LPETs themselves; observations of the reviewers using the LPET; feedback about the LPET gathered during one-on-one and group discussions; and responses to questionnaires administered before training, after training, and after several months of LPET use.

2.3. Results. As described in the sections below (and see Michael et al., 2011, for more detail), the quality reviewers who participated in the operational pilot testing found that the LPET fits easily into the operational workflow, is easy to use, and can add value to the review

process. The findings also point to some potential modifications that may improve the LPET and its training.

2.3.1. The LPET fits easily into the operational workflow and is easy to use. During testing, reviewers completed their typical review and correction processes and then used the LPET to document their review and feedback. On average, their reports of time on task when documenting their comments and feedback were the same regardless of whether they were documenting with or without an LPET, and reports from reviewers suggested that they could complete an LPET in as little as two to five minutes once they became familiar with the tool.

Reviewers shared positive comments about the LPET's clarity and ease of use, and the post-workshop questionnaires revealed positive ratings for the effectiveness of the LPET training. Reviewers generally rated the individual LPET items as intuitive and as having relatively clear distinctions between them, and they did not generally report items to be confusing or redundant.

2.3.2. The LPET can add value to the review process. One important benefit of the LPET is that it can help ensure that all reviewers are using a common set of criteria. Questionnaire data demonstrate that the LPET both captures and expands upon reviewers' pre-existing criteria regarding product quality. When reviewers were asked before training to list the characteristics of a good language product, 84% of the listed characteristics referred to properties that map closely to concepts from the LPET. Reviewers were most likely to list characteristics related to the concepts of accuracy, writing, and format, and much less likely to list features related to the dimensions of *Coverage* and *Context*. This discrepancy possibly reflects the finding that the reviewers who participated in the pilot testing typically work with full translations rather than summaries and that their customers do not generally expect them to add intelligence analysis to their language products. The LPET therefore helped these reviewers to think more broadly about the characteristics they should be looking for when evaluating certain types of language products. Reviewers also reported that the LPET increased their awareness of the relationship between source characteristics and product quality.

The LPET can also play an important role in improving feedback for language analysts. At the start of the operational testing, reviewers said that they expected the LPET to improve how consistent, systematic, and detailed their feedback for language analysts would be. After

several months of LPET use, questionnaire data confirmed that reviewers found their feedback to be more objective, comprehensive, detailed, structured, and useful than the feedback they provided without the LPET. Reviewers also believed that the LPET could help them negotiate some of the tensions that are inherent to the review process, including a sensitivity to evaluating and being evaluated. A common standard of evaluation that is applied regularly and consistently can diminish the concern that evaluations are subjective and motivated by personality rather than by product quality.

Several reviewers reported that they would recommend the LPET to others, which suggests that it helps them to do their jobs. Despite the mostly favorable responses to the LPET, however, some reviewers had critiques and suggestions for improving the tool. The reviewers who were less enthusiastic were those who already had well-established review processes. Interestingly, their recommendations often conflicted with one another. For example, some reviewers praised the comprehensiveness of the LPET, some found it too detailed, and some found it not detailed enough. The reviewers who called for more detail on the LPET were typically experienced reviewers with established methods for documenting detailed feedback for language analysts.

Because the LPET has been designed to be applied to many different product types according to the needs of diverse organizations, it is not surprising that not everyone in every organization values all of the LPET items equally or finds them sufficiently detailed. It is not realistic to expect a single tool to satisfy all needs in all situations; rather, the goal of the LPET is to provide an "80% solution" that is likely to satisfy the majority of users and their needs most of the time without sacrificing the LPET's usability, sensitivity, validity, or reliability.

2.3.3. Enhanced LPET training may help achieve valid and reliable use. Observations of quality reviewers using the LPET suggest that some aspects of the tool were not always being used appropriately. For example, some reviewers commented on major problems in the language products being reviewed while assigning fairly high ratings, whereas others made only minor stylistic changes and assigned surprisingly low ratings. Similarly, some of the completed LPETs contained a checked box for an area needing improvement, but the corresponding rating scale indicated *NA* or gave an inappropriately high rating. The LPET has since been modified to

include a series of consistency checks in the form of pop-up messages that prompt users to reconsider low or high ratings that conflict with their treatment of the checkbox items.

Although this type of consistency check will help avoid errors arising from unintentional mouse clicks, the data also indicate that some improvements may be needed to the LPET training to maximize the likelihood of valid LPET use in all contexts. For example, future training should more directly address the threshold for marking checkboxes. For the LPET to be used reliably, reviewers must agree as to whether a box should be checked any time an issue is noted, regardless of how minor the issue might be, or whether the box should only be checked to indicate a relatively serious problem.

It is also important to ensure that workshop participants have a thorough understanding of all of the items on the LPET, especially given that some LPET concepts may be unfamiliar to some reviewers depending on their experience and work environment. For the pilot participants the least intuitive items on the LPET were those related to *Context*, so it may prove helpful to devote more workshop time to the items in that dimension.

3. Next Steps

Before the LPET can be used on a large scale, it is critical to demonstrate that the tool is sensitive, valid, and reliable. We are currently developing a set of controlled experiments to test those aspects with experienced quality reviewers from throughout the US government. In the first experiment, designed to assess the validity and reliability of the product assessment checkbox items, a group of reviewers will classify errors that experts have predetermined to fit different checkbox categories. In the second experiment, a separate group of reviewers will assign ratings to language products containing sets of errors that experts have predetermined to fall within a given dimension and severity level.

By experimentally controlling the quality of the products being evaluated, we will be able to determine whether participants are using the LPET in meaningful and appropriate ways and we will be better able to determine which aspects of the LPET may need to be revised or trained differently. In addition, by collecting data from multiple reviewers for the same products we will be able to evaluate inter-rater reliability and take the next steps toward ensuring that reviewers with diverse experiences are able to use the LPET in similar and consistent ways.

To ensure that the LPET is maximally beneficial in the workplace, it is critical that users understand how to interpret LPET data. In parallel with conducting the experiments described above and implementing improvements to the LPET and training, the team is developing methods of depicting aggregated LPET data to facilitate clear interpretation and decision making. In the next section we provide some examples of the types of displays that may prove particularly useful for managers and other decision-makers.

4. Sample Views of Aggregated LPET Data

The 306 completed LPETs from the operational pilot testing provided a rich data source for developing a variety of possible display options to help users make the most of their LPET data. Figure 2 shows an example of a menu-based graphical display created with SAP Crystal Dashboard Design. In this display the average ratings on each dimension are dynamically updated based on selected values of three variables: the product type (gist, full translation, etc.), the language of the source material, and the language level (i.e., passage level) of the source material (1, 1+, 2, 2+, 3, 3+, 4). The dynamic nature of the interface allows users to view rating data at varying levels of specificity depending on their particular needs.

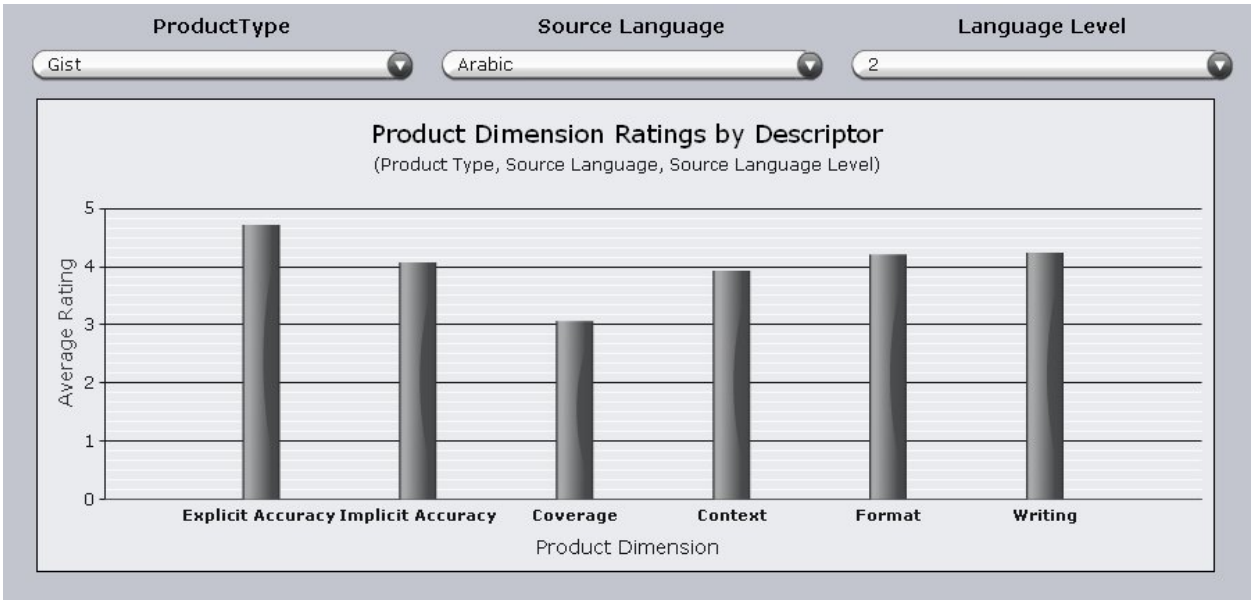

Figure 2. The gists that were generated from Level 2 Arabic source texts were given high ratings for *Accuracy of Explicit Content* but relatively low ratings for *Coverage*.

The variability across dimensions seen in this view could reflect a team's strengths and weaknesses when working with different combinations of variables. For example, the team that generated these data appears to be strongest in *Accuracy of Explicit Content* and weakest in *Coverage* when creating gists from Level 2 source material. The low ratings for *Coverage* might suggest that language analysts could benefit from additional training to help them determine what information should be included and excluded in their language products.

Managers might also compare this view with ratings for gists generated from Level 3 source materials, or with ratings of different product types. Such information could help managers decide whether particular source items or product types should be routed to a different team.

Other graph formats might be used to assess data against some benchmark. Figure 3 illustrates a combination graph in which the average ratings of the group are represented as a fixed line. The user can then select an individual language analyst (e.g., 12111) to see how his or her ratings (averaged across all of his or her products that have been reviewed with the LPET), compare to the group average.

The display in this figure could also be modified to allow for the selection of different benchmark lines; the user could thus make comparisons with different teams in the organization or with an individual language analyst's own averages from a preceding evaluation period.

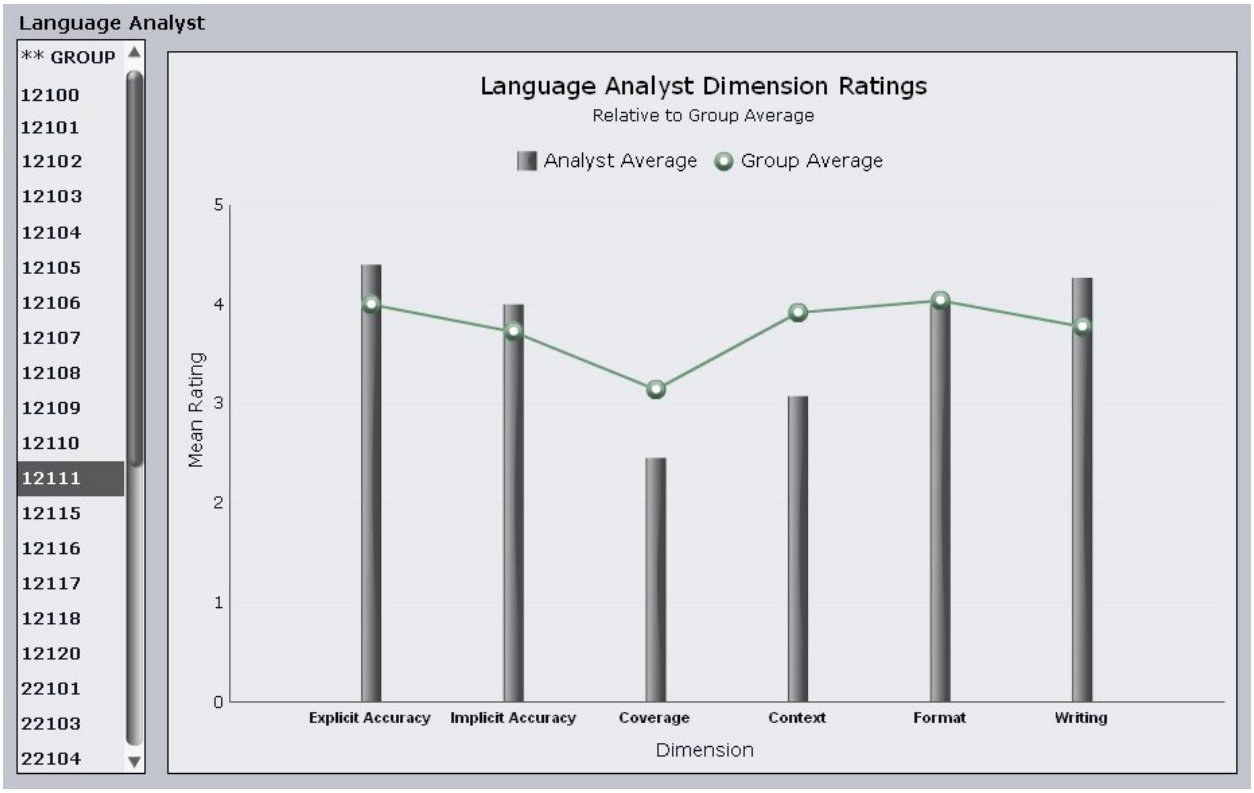

Figure 3. Language products created by Language Analyst 12111 were rated higher than the group average for *Accuracy of Explicit Content*, *Accuracy of Implicit Content*, and *Writing*. In contrast, the *Coverage* and *Context* ratings for this individual's language products were below the group average, and the *Format* ratings were similar to the group average.

5. Conclusions

The results of pilot testing the LPET in authentic work settings suggest that the LPET fits easily into the operational workflow, is easy to use, and can add value to the quality review process. Some reviewers struggled with particular aspects of appropriate LPET use, leading to minor modifications to the LPET and recommendations for enhancing the training by targeting observed areas of difficulty. Further studies are planned to evaluate the sensitivity, validity, and reliability of the LPET in controlled settings.

By providing standards for a variety of types of language products, the LPET can help ensure quality; promote consistent, systematic feedback to language analysts; provide information for planning professional development activities; and provide aggregated data to describe individual and organizational capability and performance.

6. References

Colina, S. (2008). Translation quality evaluation: Empirical evidence for a functionalist approach. *The Translator*, *14*(1), 97-134.

Colina, S. (2009). Further evidence for a functionalist approach to translation quality evaluation. *Target*, *21*(2), 235-264.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Thomson Wadsworth.

Michael, E. B., Bailey, B., Gannon-Kurowski, S., & Pinckney, K. (2007). *Description of summary translation task requirements for specific jobs and the uses of the summaries* (TTO 041 Technical Report M.19). College Park, MD: University of Maryland Center for Advanced Study of Language.

Michael, E.B., Blodgett, A., Massaro, D., Bailey, B., de Terra, D., Lutz, A., Rhoad, K., Saner, L., Gannon-Kurowski, S., Castle, S., & Pinckney, K. (2010). *Toward a standard for targeted summary translation: A test of the sensitivity, validity, reliability, and usability of the Summary Translation Evaluation Tool (STET)* (TTO 3441 Technical Report E.3.3). College Park, MD: University of Maryland Center for Advanced Study of Language.

Michael, E. B., Massaro, D., & Perlman, M. (2009). What's the bottom line? Development of and potential uses for the Summary Translation Evaluation Tool (STET). *The Next Wave*, *18*, 42-49.

Michael, E. B., Pinckney, K., Bailey, B., Lutz, A., de Terra, D., Massaro, D., & Clausner, T. (2008). *Toward a model of expert summary translation: An investigation of procedures and performance in summary translations produced under time pressure* (TTO 3441 Technical Report E.3.1). College Park, MD: University of Maryland Center for Advanced Study of Language.

Michael, E. B., Saner, L., Blodgett, A., Bailey, B., Rhoad, K., & Castle, S. (2011). *Toward a standard for language products: Pilot tests of the Language Product Evaluation Tool (LPET) in operational environments* (Project 3441 Technical Report 1.1). College Park, MD: University of Maryland Center for Advanced Study of Language.