

## **Testing the right skill: Evidence to support testing translation ability**

**Rachel L. Brooks**

Federal Bureau of Investigation, United States  
rlbrooks@fbi.gov

**Maria M. Brau**

mmbrau@gmail.com

*In the United States Government (USG), language-test scores in the various skills are defined by reference to the Interagency Language Roundtable (ILR) Skill Level Descriptions (SLDs). The ILR SLDs for Translation Performance state that reading the source language (SL) and writing the target language (TL) are prerequisites for translating but add that there is a prerequisite third skill, termed “congruity judgement”. Since the Language Testing and Assessment Unit (LTAU) of the Federal Bureau of Investigation (FBI) uses the ILR as the basis for its translation testing, it has conducted several studies on test results to discover the interconnection, if any, of these requisites.*

### **1. Introduction**

USG agencies often use a measure of reading comprehension in their foreign language to qualify a person for translation work. This practice is based on two assumptions: that as a citizen of the United States, the examinee has a sufficiently strong proficiency in English writing; and that this assumed English proficiency in combination with a measure of reading comprehension in a foreign language constitutes sufficient qualification to perform translation tasks. However, the Interagency Language Roundtable (ILR) Skill Level Descriptions (SLDs) for Translation Performance state that reading comprehension of the source language (SL) and the ability to write the TL are prerequisite but not sufficient for translation. The SLDs posit that translation also involves “congruity judgement”, defined as the ability to choose equivalent expressions in the TL that best match the meaning intended in the SL (ILR, 2005). Still, there has been little research into the relationship between reading ability and translation ability that prove the need for this additional ability (congruity judgement) and therefore the need to measure it. Past research has shown that for Arabic there is a moderate to weak correlation between reading proficiency and translation performance (Brau & Lunde, 2005), but that research was limited to one language and did not consider the demographic variables of the examinees. In addition, Howard (2016) studied translation passage difficulty, but examined it in relation to reading text difficulty, not translation difficulty. Therefore, the answer to the research question “is reading comprehension a good predictor of translation ability” remains unclear.

### **2. Background: The ILR**

In 1952, the USG Civil Service Commission called for a system to replace academic grades and self-reports with uniform language standards when hiring and evaluating employees. The Foreign Service Institute (FSI) met the challenge by convening representatives from different agencies to assist in developing a language proficiency scale, criterion-referenced testing and standardized rating factors. The FSI initiative led to proficiency descriptors and interview testing as the means of measuring speaking skills. Both the scale and the Oral Proficiency Interview (OPI) became widely accepted not only by USG agencies but also by academia (Frith, 1975; Herzog, 2003; Spolsky, 1995).

In view of these developments, and to encourage further cooperation between the agencies, the General Accounting Office in 1972 recommended a more formal organization of USG agencies. The group of agencies complied: it adopted the name ILR, agreed to operate by consensus and resolved not to seek funding for its operations from any one agency to ensure impartiality.

By 1968, several USG agencies began applying the speaking scale to other language skills, drafting a set of descriptions that not only revised the original FSI speaking scale but also included separate scales for listening, reading and writing. In 1985, the Office of Personnel Management approved these SLDs as the official standard for measuring language proficiency in the USG. From then on, hiring, promotions, duty assignments and awards for language personnel would depend on ILR scores.

By then, the Defence Language Institute (DLI) had been developing the Defence Language Proficiency Test (DLPT). Consisting of two separate parts (Listening and Reading), it was intended for learners of a foreign language whose native language was English, and was originally used to test DLI trainees. But by the late 1980s, in the absence of ILR SLDs for translation and therefore ILR-based translation tests, the DLPT became a tool for identifying prospective translators. Since reading proficiency in the SL is admittedly a *sine qua non* for translation, and since the examinees were normally able to write in English, a high score in the DLPT-Reading was considered a predictor of successful translation performance. Likewise, an OPI was increasingly considered sufficient for choosing an interpreter, particularly since interpretation was listed as one of the tasks for a Level 4 speaker (Child, 1987).

How adequate these instruments were in assessing translation and interpreting skills became a subject of discussion in the USG, and as early as the late 1980s some agency representatives were arguing that the “combined skills” merited separate consideration. In a volume on translation issues, sponsored by the American Translators Association (ATA), the term “congruity judgement” was introduced by the Director of Language Testing at the Department of Defence (DOD) to describe the “x” factor that enabled a translator to capture and convey the full meaning of the ST (Child, 1987).

As a result, by the early 1990s, representatives from some agencies were holding their own meetings, separate from the ILR, to discuss and attempt to resolve translation and interpreting issues. When the ILR did not follow with any action, the Federal Bureau of Investigation (FBI) produced guidelines for its own use that were not ILR-approved and were therefore labelled “Provisional Interagency Language Roundtable Skill Level Descriptions for Translation”. Likewise, a separate effort on the part of DOD personnel produced a set labelled “ILR SLD for Translation (Congruity Judgement)” that were published as such without having been presented to the ILR for approval (Cascallar, Cascallar, Lowe and Child 1996, Stansfield, Scott and Kenyon, 1992). Based on these Provisional Descriptions, the FBI proceeded to develop a series of translation tests.

Because language acquisition was its focus, the ILR had initially shown little interest in dealing with the assessment of translation and interpreting, and functioned with two working committees only: Training and Testing. However, in 1995 the ILR finally organized a Special Interest Group (SIG) to accommodate those members who had been discussing the so-called combined skills. These matters stood until the demand for translation and interpreting services surged after 11 September 2001. At that point, the SIG was promptly elevated to committee status on a par with the two original Testing and Training committees. The new Translation and Interpretation (T&I) Committee then moved to meet the needs of the USG. By 2003 it had produced SLDs for Translation Performance, which were approved by the ILR in 2004. SLDs for Interpretation Performance were similarly developed and approved by the ILR in 2007, followed by SLDs for Audio Translation in 2012.

For the ILR, these new descriptions represented a significant shift in focus from proficiency to performance. Furthermore, they were not pegged to the progression of language learners along a scale, a process clearly apparent in the original SLDs, while at the same time “native speakers” no longer constituted a frame of reference.

Intended for USG managers, the Preface to the ILR Translation Performance SLDs (2005) clearly stated:

Competence in two languages is necessary but not sufficient for any translation task. Though the translator must be able to (1) read and comprehend the source language and (2) write comprehensibly in the target language, the translator must also be able to (3) choose the equivalent expression in the target language that both fully conveys and best matches the meaning intended in the source language (referred to as congruity judgement). (lines 16-20)

A weakness in any of these three abilities will influence performance adversely and have a negative impact on the utility of the product. Therefore, all three abilities must be considered when assessing translation skills.

The prerequisite language skills for translation (reading in the source language and writing in the target language) are tested separately from translation skills themselves. Language proficiency testing serves as a screening tool and may provide a preliminary indication of the individual's potential as a translator. However, to assess translation skills, a translation test that measures the individual's ability to exercise congruity judgment and apply a translation methodology successfully must be used (ILR, 2005).

Based on the new ILR Translation SLDs, the FBI proceeded to develop a set of Translation Examinations in 30 languages (Brau, 2013). The examinations provide passages of increasing translation difficulty level as defined by the ILR, in terms of how often congruity judgement must be exercised in order to produce an acceptable translation.

However, reading comprehension may still be considered as a determinant of text typology. In a recent study, Howard (2016) uses the ILR Reading SLDs instead of the ILR SLDs for Translation Performance as the basis for determining text difficulty, therefore linking reading text difficulty to translation text difficulty.

### **3. FBI studies**

Linguists at many USG agencies were trained in government language schools. Since these language learners were considered to have native English proficiency and could write English well, it was presumed that adding a foreign-language reading skill would qualify them as translators. By way of contrast, the FBI, which does not have a language school, found itself relying largely on linguists whose native language was the foreign language, for whom reading comprehension of the source was generally not a challenge. This, plus the recognition that translation was more than reading plus writing abilities, led the FBI to develop translation tests for linguist qualifications.

In the aftermath of the 9/11 attacks, the FBI received thousands of applications for the position of Arabic language analyst. At the time, the complete test battery consisted of five tests: the DLPT Listening/Reading as a screening tool, a Translation Test (TT) from the foreign language into English, and two Speaking Tests (one in English, the other in the Foreign Language). The TT had been developed based on the Provisional Descriptions produced in-house.

Because of the time and effort involved in rating TTs and the need to speed up the process, it was suggested that the FBI follow the example of other agencies and use the multiple-choice DLPT Reading as a predictor of translation performance, thereby delaying or even eliminating the TT. To justify using a testing instrument that assessed translation, that is, congruity judgement, as the predictor, the FBI conducted a series of three studies.

#### **3.1 Reading and translation**

In 2005, Brau and Lunde presented a study posing two research questions, the first being: Is reading comprehension a good predictor of translation performance? According to the accepted practice of using reading scores to assign translation work, scoring well on reading should correlate with scoring well in a translation examination. However, if reading did not prove to correlate with translation, it would be important to find out what did. Therefore, the second question was posed. What other factors contribute to translation ability?

The dataset for the study came from the pool of Arabic linguist applicants who had applied in the four months after the 9/11 attacks, when then Director Mueller made a call to American citizens to join the FBI and use their language skills to help fight terror. Thousands applied for the position; they had varying backgrounds and levels of experience in translation. Of these applicants, there were 1,438 who passed the pre-screening security requirements and began the testing process. Of the examinees, 63% reported their primary language as being Arabic; 23% reported it as being English; 10% claimed both Arabic and English as their primary languages, leaving 4% declaring other languages as their primary language.

The applicants had to pass a battery of five tests to qualify for the position. Phase 1 testing included the Defence Language Proficiency Tests (DLPTs) for Reading and Listening in Modern Standard Arabic as well as the TT from Modern Standard Arabic into English. The DLPTs were multiple-choice tests with a raw score (0–60) that converted to an ILR rating, with Level 3 as the highest possible score. The DLPTs were used as a filter, with ILR Level 2+ (Limited Working Proficiency, Plus) as the score that qualified applicants to continue testing for a variety of language positions. The TT was rated by trained, experienced raters. The TT raters assigned an expression rating which reflected the examinee’s English expression, and an accuracy rating, which reflected how accurately the source had been conveyed. The lower of the two became the final rating. Once an applicant had successfully completed Phase 1, a full background investigation was initiated, eliminating approximately two-thirds of the examinees. Only 99 applicants remained eligible to complete Phase 2 testing, which included Speaking Proficiency Tests (SPTs) for both Modern Standard Arabic and English.

Test results from Phase 1 (Table 1) indicated that slightly more than half of the examinees (52.5%) passed the DLPT Listening examination at a minimum of ILR Level 2+. A larger proportion of the examinees, equivalent to 74.2%, passed the DLPT Reading examination at ILR Level 2+ or higher. Approximately 20% of the examinees passed the reading and the translation tests at an ILR level 2+ or higher. Slightly more than half of all the examinees passed the reading test, but failed the TT (54.3%). Only 289 examinees (20.1%) passed the TT, making it the most discriminatory test in the battery. Less than 1% of the examinees failed the reading test at ILR Level 2+ but did pass the TT, and all of these failed the reading test by only one point.

Table 1: Phase 1 testing results

	Total
Total applicants	1,438
Passed listening	755 (52.5%)
Failed listening	683 (47.5%)
Passed reading	1,067 (74.2%)
Failed reading	371 (25.8%)
Passed TT	289 (20.1%)
Failed TT	1,149 (79.9%)
Passed reading/failed TT	781 (54.3%)
Failed reading/passed TT	12 (0.8%)

The background examination between Phases 1 and 2 whittled down the examinee pool from the 286 who had passed Phase 1 to only 99 who went on to complete both speaking tests in Phase 2. The Arabic and English speaking tests each had a 94% pass rate at ILR Level 2+ or higher.

As the research questions indicate, the study aimed to explore whether a measure of reading is sufficient substitute for a measure of translation. Considering that the demographics of the FBI

applicant pool included mostly native and heritage speakers of Arabic, the researchers hypothesize that examinees' reading ability in their self-reported primary language (Arabic) would not be as strong a determining factor as their ability to express themselves in English. Therefore, correlations between the English expression rating from the TT were also compared to the final rating of the TT. Initially, Pearson correlations were run between the DLPT-Reading raw score, as it revealed more granularity, and the final rating of the TT for both all the examinees and those who reported both Arabic and English as their primary language (Table 2). The correlation for all three examinee groups was fairly strong and significant. The correlation between reading and translation for all examinees was .70, with a weaker correlation for the Arabic primary language group (.67) and a stronger correlation for the English primary language group (.76). Even stronger and still significant correlations were found when comparing the English expression ratings of the TT to the overall TT rating. The correlation for all the examinees was .75, with a stronger correlation for the Arabic primary language group (.78) and a weaker correlation for the English primary language group (.74).

Table 2: Pearson correlations of translation compared to reading and expression scores

Primary language	All applicants n = 1,438		Passed reading/ failed TT n = 781		Passed TT n = 289	
	Reading/ final TT	Expression/ final TT	Reading/ final TT	Expression/ final TT	Reading/ final TT	Expression/ final TT
All examinees	.70*	.75*	.35	.51	.11	.40
Arabic	.67*	.78*	.37	.53	.06	.35
English	.76*	.74*	.25	.46	.16	.40

\* Correlation is significant at the .01 level (2-tailed)

A visual examination of the raw data revealed that much of the strength in the correlations was due to the fact that a low reading score correlated very strongly with a low translation score, which confirms the notion that reading is a prerequisite for translation. However, the question of interest for hiring and assignment decisions is whether a high reading score correlates to high translation ability. To investigate this issue, a subset of the examinees was analysed: those who passed the DLPT-Reading but failed the TT (n = 781). The comparison between the reading and the translation scores revealed a relatively low and non-significant correlation overall, with a slightly stronger correlation for the Arabic primary language group (.37) and a weaker correlation for the English primary language group (.25). For this subset, it was the case that the English expression scores of all primary language groups showed a stronger but not significant correlation to the translation score: for all the examinees (.51), for the Arabic primary language group (.53), and for the English primary language group (.46).

Importantly, the researchers were interested in seeing which other abilities correlated well with those who passed the TT, so that examinee group (n = 289) was analysed separately. Here it is apparent that the reading score has a low (and not-significant) correlation with the translation score for all primary language groups: for all the examinees (.11), for the Arabic primary language group (.06), and for the English primary language group (.16). The English expression score from the TT compared to the final TT score resulted in a much better, but still only moderate, correlation: for all the examinees (.40), for the Arabic primary language group (.35), and for the English primary language group (.40).

In conclusion, the data support the assertion that reading ability in the SL and writing ability in the TL are prerequisites for translation ability, but they are not substitutes for measures of translation ability. The inability to read the SL or write the TL will predict low translation ability; however, high reading and writing abilities also cannot predict translation ability. Therefore, the ability to exercise congruity judgement is a latent variable that plays an integral role.

This study did have two important limitations. First, all the subjects were Arabic applicants and results may vary depending on language. The study needed to be repeated with additional languages. Second, the expression score was a sub-score in the TT and was not a measure of free expression, but a measure of expression while translating. This rating was therefore not equivalent to an independent rating of English writing proficiency.

### 3.2. Writing and translation

The FBI pursued additional research to overcome the language and written expression measurement limitations of Brau and Lunde (2005). In Brooks and Brau (2006), they presented a study that investigated other predictors of translation ability. Two research questions were proposed: first: Is English writing ability a good predictor of translation ability for those whose native or primary language is the SL? Since most FBI linguist applicants are native and heritage speakers of the foreign language (the SL for translation texts), further exploration of the weaker of the two prerequisite skills (reading the source text and writing the TL) was warranted. Additional information was available on the examinees' self-assessments, so the question: "Is self-assessment a useful predictor?" was also posed.

The data for this next research project were taken from the validation study of the new translation tests that the FBI developed, the Verbatim Translation Examinations (VTEs). Three of the 15 languages in which the new tests were developed were selected from the study: Italian, Vietnamese and Turkish. Each language group had 80 examinees, giving a total of 240 participants. They were all at least 18 years of age and were selected to represent FBI's linguist applicant population, approximately 90% native and heritage speakers. As with most FBI applicants, they had various levels of higher education.

The concordances of ratings listed in Table 3 show the percentage of matches between the translation rating (VTE) and the English writing proficiency rating (EWT) or the self-assessment. All three languages combined show that there is only an exact match between translation ability and English writing ability in 8.69% of the cases among the three languages. Considering that 90% of the participants in the study were native and heritage speakers and that English writing would be their weaker prerequisite skill, this concordance is very low. It is interesting to note that Vietnamese has a higher exact match concordance at 19.82% than Italian (3.75%) and Turkish (2.50%). There is a much higher percentage of cases where the translation scores and the English writing scores differed by only a plus level but were within the same range (35.07%) or were one whole ILR Level different from each other (24.59%). It is interesting to note that the writing proficiency assessment is never (0.00%) more than a level lower than the translation score in any language case.

Table 3: Concordances between translation ratings of English writing scores/self-assessments

		Italian overall (%)	Vietnamese overall (%)	Turkish overall (%)	Average overall *%)
Exact match	VTE and EWT score	3.75	19.82	2.50	8.69
	VTE and self-assessment rating	10.46	26.11	13.90	16.82
Within level match	VTE and EWT score	28.75	50.16	26.30	35.07
	VTE and self-assessment rating	20.65	19.76	15.20	18.54
One level difference (VTE higher or lower)	VTE and EWT score	32.50	12.57	28.70	24.59
	VTE and self-assessment rating	20.81	30.92	27.90	26.54
Greater than one level difference (VTE higher)	VTE and EWT score	0.00	0.00	0.00	0.00
	VTE and self-assessment rating	0.00	0.00	0.00	0.00
Greater than one level difference (VTE lower)	VTE and EWT score	35.00	17.45	42.50	31.65
	VTE and self-assessment rating	48.08	23.21	43.00	38.10

VTE: Verbatim Translation Exam

EWT: English Writing Test

It appears that the English writing ability is either the same as the translation ability or higher, giving weight to the argument that writing the SL is a prerequisite of the ability to translate. Even though the translation and writing final ratings are close in many instances, it is still important to note that the translation rating is more than a full ILR level lower in almost one-third of the cases (31.65%). Agreement between the two abilities overall seems best for Vietnamese and worst for Turkish: translation ability for the Turkish group was more than a level lower than English writing ability in 42.5% of cases.

Interestingly, the self-assessment is a stronger predictor of translation ability than English writing. The self-assessment rating was an exact match overall with the translation rating in 16.2% of cases, twice as much as the English Writing Test but still a low predictor. In 38.10% of cases, the participants overestimated their translation ability by more than a whole ILR level.

When Pearson correlations were run for each of the language groups, the results aligned well with what was seen from the concordances (Table 4). Vietnamese has a moderately strong correlation (.62) between English writing and translation abilities, but not between translation and the self-assessment (.33). On the other hand, there was a stronger correlation between the self-assessment and translation ability (.40) than between the translation test and writing ability (.27) for the Italian participants. It should be noted that these two correlations are fairly low and would not make a good replacement for a TT. The Turkish participants also had a moderate correlation between writing and translation (.46). The correlation between the self-assessment and translation ability was not very strong (.40), but similar to that of the other two language groups.

Table 4: Pearson correlations of translation compared to writing and self-assessment scores

	Italian	Vietnamese	Turkish
	<i>Correlation</i>	<i>Correlation</i>	<i>Correlation</i>
VTE and EWT	.27*	.62**	.46**
VTE and Self-Assessment	.40**	.33**	.40**

\* *Correlation is significant at the .05 level (2-tailed).*

\*\* *Correlation is significant at the .01 level (2-tailed).*

In conclusion, neither writing proficiency assessments nor self-assessments are an accurate or reliable predictor of translation ability. This holds true even when the participant population is mostly heritage and native speakers, which would presumably mean that their reading comprehension of the SL would be stronger than their ability to produce the TL. Translation ability must therefore be tested as a separate skill and not derived from measures of reading the SL and writing the TL.

### 3.3 Reading and writing together with translation

Lunde and Brau (2005) and Brooks and Brau (2006) began the work of exploring the role of reading and writing proficiencies in translation performance, but they looked at these skills individually compared to translation and have limitations in the number of languages analysed. By 2015, VTEs in more than 20 languages had been in use for approximately ten years in an authentic context (rather than a validation project with paid participants) and a large amount of data had been collected. The FBI updated the previous research by adding more examinees and a broader range of tests. It provided a better opportunity to test the roles of reading and writing in translation with the new VTE, which is ILR-based.

Data over ten years of applicant processing were collected; it comprised 10,875 linguist applicants who had taken the VTE. Examinees with incomplete test batteries and languages that did not have more than 40 examinees were purged from the data. The total number of examinees that had reading test scores and VTEs was 7,524 and those who had reading scores and writing expression scores (a sub-score of the VTE) and VTE scores numbered 4168. ILR plus levels were interpreted as 0.6, as a rating of 1+ is represented as 1.6.

A review of the descriptive data from the study, shown in Table 5, reveals that the average VTE ratings by language range from 1.36 (Japanese) to 2.14 (German), with an average across all languages of 1.8. The average DLPT rating across all languages is 2.84, more than an entire ILR level higher than that for translation. Language for reading range averages from 2.55 (Bosnian) to 2.95 (Bulgarian). The average English expression rating (2.45), a separate sub-rating of English writing proficiency from the VTE, is also higher than the overall VTE average by an ILR sub-level (0.65).



Table 5: Means and SDs for test ratings by language, sorted alphabetically

Language	VTE mean	VTE SD	DLPT-R mean	DLPT-R SD	Eng exp mean	Eng exp SD	DLPT minus VTE	EXP minus VTE
Arabic	1.82	0.45	2.63	0.42	2.38	0.52	0.81	0.56
Bosnian	1.79	0.47	2.55	0.76	2.35	0.41	0.76	0.57
Bulgarian	2.00	0.53	2.96	0.16	2.57	0.57	0.96	0.57
Chinese	1.81	0.49	2.89	0.31	2.42	0.52	1.08	0.61
Farsi	1.68	0.52	2.87	0.40	2.30	0.53	1.19	0.63
French	1.78	0.47	2.81	0.34	2.71	0.55	1.02	0.92
German	2.14	0.55	2.95	0.20	2.79	0.68	0.81	0.65
Hebrew	1.97	0.46	2.89	0.35	2.55	0.56	0.91	0.58
Italian	1.70	0.34	2.80	0.35	2.52	0.53	1.10	0.82
Japanese	1.36	0.53	2.80	0.33	2.35	0.62	1.44	0.99
Korean	1.53	0.53	2.92	0.22	2.22	0.66	1.39	0.69
Portuguese	1.75	0.37	2.80	0.34	N/A	N/A	1.05	N/A
Russian	1.92	0.57	2.89	0.23	2.47	0.61	0.97	0.54
Turkish	1.75	0.49	2.91	0.29	2.27	0.60	1.16	0.52
Vietnamese	1.86	0.57	2.89	0.29	2.29	0.62	1.03	0.43
MEAN	1.80	0.52	2.84	0.35	2.45	0.58	1.04	0.65

The reading comprehension score was compared with the translation rating independently at first. Out of 7,524 examinees, only 34 had translation scores higher than their reading scores, or 0.5% of ratings. There were 113 instances where the TT score was much lower (more than two whole levels) than the translation rating, indicating that a high reading rating does not mean a high translation rating. The English expression ratings were also compared with the translation ratings for the 4,168 examinees that had both. Only 0.2% of these examinees had translation test ratings higher than the English expression ratings. There were 173 instances where the English expression rating was two or more ILR levels higher than the translation rating.

A regression analysis was conducted to determine to what extent the two independent variables (reading proficiency and writing proficiency) predicted the dependent variable, translation performance, for the sample that had both figures ( $n = 4,168$ ). The analysis computed the following formula:

$$\text{VTE Rating} = 0.5054 * \text{Expression Rating} + 0.2004 * \text{Reading Comprehension Score}$$

This formula was significant at  $p < 0.01$ . The formula has an R statistic of 0.98, which shows a very strong positive correlation between reading, writing and translation. The adjusted  $R^2$  is 0.95, which indicates that 95% of the data fit the model. Next, the data were placed in the model to see if the fit was as good as the statistic indicated. The formula was used to calculate a predicted translation rating for each examinee. The average of these ratings is displayed in Table 6.

Table 6: Actual and predicted VTE ratings

VTE rating	Predicted rating mean	Predicted rating SD
0	0.53	0.06
0+	1.23	0.40
1	1.54	0.32
1+	1.73	0.25
2	1.91	0.22
2+	2.09	0.18
3	2.22	0.17
3+	2.43	0.10
4	2.62	0.00

The mean predicted rating was only fairly accurate for the Actual VTE rating group for Level 1+ (predicted 1.73 is not far from 1.6) and Level 2 (predicted 1.91). At the low and high ends of the VTE scores, the formula does not fit. The low end shows about a plus level difference between actual and predicted scores and the high end underestimates ability by a plus level to a whole level or more. This amount of variation would give too many false positives for the high-stakes nature of FBI testing. These instances of misfit are also clear in the regression line in Figure 1.

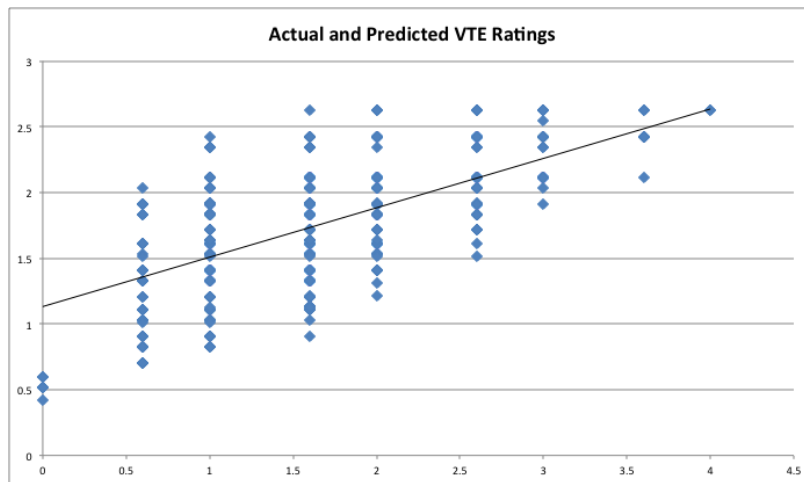


Figure 1: Actual and predicted VTE ratings with regression line  
 x-axis: actual VTE ratings  
 y-axis: predicted VTE ratings

This final and more comprehensive study confirmed the previous conclusions: measures of reading comprehension or English writing ability alone are not reliable predictors of translation ability. Reading comprehension in the source language and writing ability in the TL are components of translation and together they may give a more accurate prediction of translation than individually. Nevertheless, the predictions are not accurate enough for high-stakes decisions: an absent skill is needed to make better predictions for translation: congruity judgement. The congruity judgement

component must be measured through a translation performance test and reading and writing tests should be used only for screening purposes.

#### 4. Conclusions

Taken together, the three studies discussed have provided a thorough examination of the role that reading proficiency and writing proficiency play in translation ability. Brau and Lunde (2005) show that there is a significant correlation between reading and translation, but that that correlation is strengthened by the fact that low reading ability leads to low translation ability. When that is removed, the correlation is very weak. Brooks and Brau (2006) also show a moderate correlation between writing proficiency and translation performance, but it varies significantly from language to language. Brooks and Brau (2015) finally combine the concepts of the two previous studies with a broader range of languages to create a formula for predicting translation ability from reading and writing scores. Although the formula initially appears to have a good fit, the decisions are often off by a plus level and in many cases a full ILR level or more. This level of error is not sufficient for the high-stakes decisions at the FBI and is therefore unacceptable. It is confirmed that there is another competence other than reading and writing that must be measured.

In summary, there is sufficient data to support several propositions, all of which appear in the Preface to the ILR SLDs for Translation Performance. Fundamentally, reading comprehension in the SL and writing ability in the TL are prerequisites for translation performance. As a result, reading and writing tests are useful for screening purposes prior to a subsequent translation assessment. However, measures of reading comprehension or English writing ability alone are not reliable predictors of translation ability. Predictions based on reading scores are not sufficiently accurate for making high-stakes decisions. Consequently, congruity judgement must be measured by means of a translation performance test.

#### References

- Brau, M. M. (2013) ILR-based verbatim translation exams. In E.D. Galaczi & C. Weir (Eds.), *Exploring language frameworks: Proceedings of the ALTE Krakow Conference 2011. Studies in Language Testing 36* (pp. 333–344). Cambridge: Cambridge University Press.
- Brau, M. M., & Lunde R. M. (2005, July). *Correlation between reading and translation ability*. Paper presented at the World Congress of Applied Linguistics, Madison, WI.
- Brooks, R. L., & Brau, M. M. (2006, June). *Correlation between English writing skills and translation ability*. Paper presented at the American Association of Applied Linguistics, Montreal, Canada.
- Brooks, R. L., & Brau, M. M. (2015, March). *Testing the right skill: The misapplication of reading scores as a predictor of translation ability*. Paper presented at the Language Testing Research Colloquium, Toronto, Canada.
- Cascallar, E. Cascallar, M. I., Lowe, P., & Child, J. R. (1996). Development of new proficiency based skill level descriptions for translation: Theory and practice. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Colloquium. Studies in Language Testing 3*, (pp. 292-313). Cambridge: Cambridge University Press.
- Child, J. R. (1987). Language proficiency and translation. In M. Gaddis Rose (Ed.), *Translation excellence: Assessment, achievement, maintenance. American Translators Association Scholarly Monograph Series*, Binghamton, NY University Center at Binghamton (SUNY).
- Frith, J. R. (1975). Cooperation among governmental agencies in the language field, *ADFL Bulletin*, 6(4), 15–17.
- Herzog, M. (2003). An overview of the history of the ILR Language proficiency skill level descriptions and scale, *Interagency Language Roundtable*, Retrieved 15 July 2011 from <<http://govtilr.org/Skills/IRL%20Scale%20History.htm>>.
- Howard, D. L. (2016). *A quantitative study of translation difficulty based on an analysis of text features in Japanese-to-English short-passages translation tests* (Unpublished doctoral dissertation). Universitat Rovira i Virgili. Tarragona, Spain.

- Interagency Language Roundtable (ILR). (2005). Interagency Language Roundtable Skill Level Descriptions for Translation Performance. Retrieved 1 September 2011 from <http://govtilr.org/Skills/AdoptedILRTranslationGuidelines.htm>.
- Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.
- Stansfield, C. W., Scott, M. L., & Kenyon, D. M. (1992). The measurement of translation ability. *Modern Language Journal*, 74, 455–467.