Final Project
SLS 670
Dennis Koyama and Dan Silver

Abridged Estimate of Inter-rater Reliability in ELI Writing Assessment

Rationale for this project

There are two main reasons that we decided to undertake the current project. The main impetus was our own direct involvement as raters of the writing tests. We had read enough tests to see many instances of very high reliability (where all three raters agreed with no ambivalence about a student's placement) and some instances of high variability in raters' scoring. Perhaps because this variability was sometimes shocking when it did occur, it made us curious about the the resulting reliability (or lack of reliability) of the overall test. A second reason for our interest in completing the project arose from an early meeting with the ELI director. During this meeting, we discussed ELI writing placement testing procedures and changes that had occurred to those procedures historically. We realized that no quantitative evaluations had been performed on the new versions of the test. We felt that it would be appropriate to take a look at the reliability estimates of the new test to see if they reached levels that the ELI would consider acceptable.

It should be noted that there was some confusion over whether or not reliability estimates could be performed on the existing data. There were two issues, in fact, which may have prevented previous ELI testing personnel from running these reliability estimates. The first involved major limitations with the scale of the test (each test is ostensibly rated on a three-point scale: 73, 83, or Exempt) and a smaller but important limitation involved a lack of consistency in the rater population.

The first issue, regarding the scale of the test, was mitigated when we realized that the ratings could be coded into seven categories, rather than only three. This was possible because of the +/- scores that raters often assign to the ratings (discussed above). The presence of seven rather than three scale points allowed us to perform these measures but the scale size remains

relatively small and probably obscures these estimates somewhat. As a justification for our use of this seven-point coding system, we have included figures of frequency counts in Appendix A to demonstrate that all points on the scale were indeed used by raters.

The other limitation regarding rater consistency was caused by the randomizing of the groups of raters. That is to say, three or more readers rated every test, but these readers were not consistently the same three individuals. (Every test received three ratings, but generally there were six or more raters present at each rating session.) We chose to deal with this problem by creating three institutional rater slots in our spreadsheet. The actual raters' scores were then subsumed into these slots and in this way numbered, rather than named, institutional raters were created. These were the raters that we used to compute inter-rater reliability.

Analysis of current rating procedures

Coding

As previously mentioned, we coded scores using a seven point numerical scale (1-7).

coding	reader rating
1	73
2	73+ *or* (73) / 83
3	83- *or* 73 / (83)
4	83
5	83+ *or* (83) / EX
6	EX- *or* 83 / (EX)
7	EX

It should be clear from this figure that there was some variation in the way raters coded their intermediate placements. Some raters used a plus/minus system and other raters wrote the two placements and chose one of them by circling the value s/he was "leaning" toward. We treated these two rater code types as equivalents when we changed these values into numerical form.

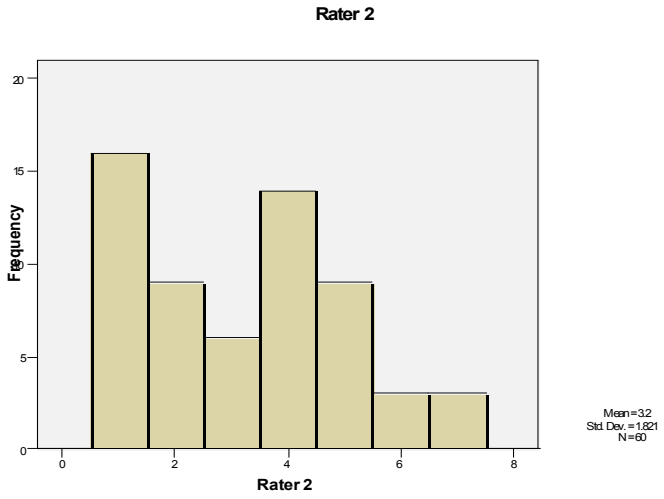
Occasionally more than three raters were used when the first three readers' ratings varied significantly. For the purpose of much of this analysis we have chosen to truncate the additional ratings so that we can look at a larger number of tests and we can ascertain the reliability of the test that obtains prior to the reliability-boosting measure of adding additional raters.

Descriptive Statistics

The table and figure (for the second rater only) below reveal the dispersion of the test ratings. Other raters show similar patterns and are located in Appendix A.

Rating Dispersion for Whole Test

	Mean	Std. Deviation	N
Rater 1	3.57	1.925	60
Rater 2	3.20	1.821	60
Rater 3	3.10	1.875	60

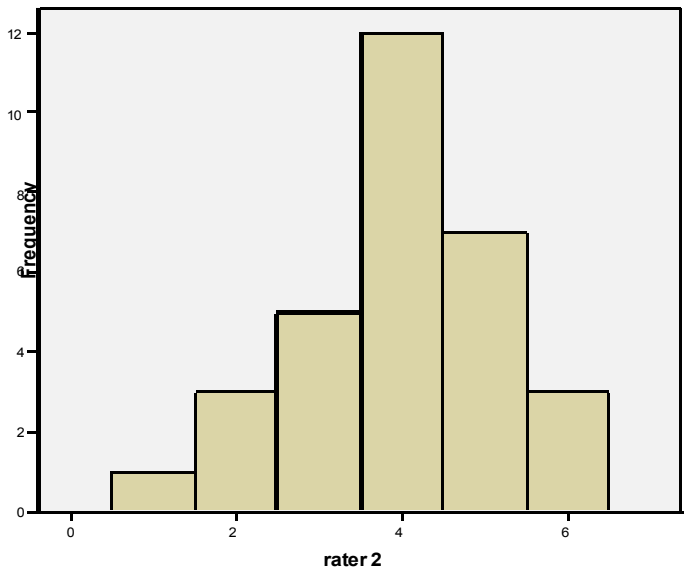


With a standard deviation of almost 2 in a 7-point scale we are clearly not looking at a normal distribution. The histogram confirms this. In fact this is no surprise since this test taken as a whole is essentially criterion-referenced, not norm-referenced. However, when we separate out only those tests whose final placement was ELI 83, we see a much more normal distribution (see

table and figure below), indicating that, perhaps for the purpose of analysis, the ratings may be more or less normally distributed when taking each placement level by itself. This is a function of the design of the test, since it is expected that ratings would cluster around the three poles. (However this normal distribution only works with the ELI 83 level since the 73 and Exempt levels will be expected to show only the top half and bottom half of normal curves, respectively.)

Rating Dispersion for ELI 83 Placement Level

	N	Mean	Std. Deviation	Variance
rater 1	31	4.55	1.234	1.523
rater 2	31	3.97	1.224	1.499
rater 3	31	3.65	1.226	1.503
Valid N (listwise)	31			



It can be observed in this histogram of rater two's ratings that 24 papers were rated ELI 83 (between points 3 and 5) and were in the final evaluation actually placed into ELI 83. On the other hand, seven papers that ended up in ELI 83 were placed either below that level or above that level in this rater's initial estimation.

Inter-rater Reliability

Inter-rater reliability was tested using two sets of measures: Pearson's *r* and Cronbach's □ These two reliability estimates have different functions: *r* is used as a correlation between only two sets of ratings while □is a reliability estimate that can be performed across more than two sets of ratings. While both estimates may be used, we prefer □over *r*, since the *r* calculations are unable to take into account the added reliability usually occurring when three or more raters are present. Having said this, administrators may find it worthwhile to observe variance between individual sets of raters, so Pearson's *r* calculations are included in additional tables in the section entitled *Additional Lines of Inquiry* along with brief commentary about their possible significance.

Inter-rater Reliability for 3 Rater Group

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.801	.802	3

Inter-rater Reliability for 4 Rater Group

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.770	.768	4

As can be noted in the tables above, a conservative estimate of ELI inter-rater reliability for the 3-rater group in the Fall semester of 2006 was right around the .80 threshold generally expected of placement tests. With more raters, higher reliability should always be expected. Therefore it would be surprising that the 4-rater group has a somewhat *lower* reliability estimate

if we didn't know that this group was, by definition, made up of papers which the first three raters had trouble agreeing on and to which a fourth rater was added. While it is always ideal to produce high inter-rater reliabilities there are potential constraints on boosting this reliability. Perhaps the biggest constraint is caused by the subject matter being tested. The more abstract the constructs being measured, the more difficult it becomes to achieve a high inter-rater reliability. The inverse is also true, for example, a mathematics exam with a concrete answer key should have an extremely high inter-rater reliability coefficient approaching 1.0. (Penny et al.) Information about constraints on inter-rater reliability for writing measures, along with recommendations related to boosting reliability is found in the recommendation section of this paper.

Standard Error of Measurement

Another useful estimate of reliability can be produced by calculating the standard error of measurement of a test. According to Brown (2005) this measure can be more useful for administrative decision-making related to the placement test since it indicates the probabilities of change in a student's placement if the student is retested using the same measure.

Standard Error of Measurement

	Three Raters	Four Raters
Whole Test	.80	1.03
ELI 83 Only	1.00	too few samples

These numbers provide a "window" around the placement scores using the proportions and assumptions of a normal distribution of measurement errors. For example, if an ELI 83 student rated by three raters receives a score of four, then there is a 68% chance[1] that the next

[1] 68% is the fixed statistical calculation of the amount of samples which occur inside the window of +/- one standard error of measurement (or one standard deviation).

time s/he is rated, his score will be between 3 and 5 (the score +/- the standard error of ELI 83 as observed in the above chart). These figures are somewhat more worrisome, in our opinion, than the reliability estimates of Cronbach's alpha found on this test, since they show a fairly high probability that from a statistical perspective, borderline placements may sometimes be arbitrarily made.

Additional Lines of Inquiry

One other brief analysis was performed on the data. Correlation coefficients were obtained between each individual rater (actual individuals, unlike the institutional raters discussed above) and the final placement. A cursory, visual examination seemed to reveal a difference between native and non-native speakers and a difference between writing experts and non-writing experts. However, high correlations between a particular rater's scores and the final placement could be rooted in a variety of causes. For example, in our experience, non-writing teachers and non-native speakers sometimes defer to the judgments of those whom they see as more knowledgeable than themselves. Or, conversely, writing teachers may have stronger convictions about a particular placement since they are more familiar with course objectives and have a better idea of the suitability of a student for a particular course. The strength of this conviction could then influence the placement outcome during the discussion part of the placement process. These are just two possible interpretations for these preliminary results. However, we do not recommend action based on these results since they compose a small n-size (less than 10). This line of inquiry could be more fully developed using both qualitative and quantitative methodology. In any case, creating some kind of a homogenous group of raters could serve to boost the reliability of this test, although this may not be considered acceptable or practical administratively.

Correlation of 10 raters with final placement, according to rater-type

Native	Non-native	Writing	Non-writing
.81	.75	.79	.81
.79	.64	.79	.64
.79	.62	.75	.62
.73	.60	.73	.60
.67	.44	.67	.44

Recommendations

Recommendation #1

One of the major limitations to the data presented in this paper is the frequent presence of rater coding inconsistencies. We found several instances of ambiguous placements, such as 73 / 83 (neither of them being circled). Another frustrating component was the high variability between the individual rater sheets and the master (orally reported) rater sheet. Since each rater first writes down scores on his/her own sheet and then during the discussion time those scores are orally reported, we hypothesize that part of the inconsistency may be coming from the orally reporting. More than thirty out of sixty sets of scores on the master sheet were somewhat inconsistent with what was found on the individual raters' sheets. The most benign interpretation of this is that raters may be changing their scores in their heads after they write them down. Other possibilities are that the leader of the rating session is not coding the scores accurately on the master sheet (perhaps writing simply 73 in place of an orally reported 73+) or the raters, for social reasons adhere to the mean and "soften" more extreme scores when they report them orally. In any case, this problem makes it difficult to check reliability. We dealt with it by basing our analyses on the scores found on individual raters' sheets. Where there were discrepancies or missing data, we used the master sheet as a supplement. Giving directions to raters about the need for their rating to be specific and consistently reported (oral reporting should be the same as written reporting and if oral reporting changes, the written report should also change) would help

clean up this data for researchers who want to look at reliability in future semesters. This change is highly recommended.

Recommendation #2

When three raters produce diverse scores, the ELI should consider adding two additional raters instead of only one. Many of the ratings are very reliable, but these unreliable sets of three are sometimes so varied that it may be possible that the final rater is a de facto solo rater. For example, if the three previous raters give scores of 1, 4, and 7 (73, 83, Exempt) these ratings appear random enough to have the effect of canceling each other out. The final rater may produce the final placement by him/herself in this case. The addition of a fifth reader in these cases will help to identify emerging patterns of placement and take some of the weight off of rater four's shoulders.

Recommendation #3

We recommend that the ELI consider editing the hallmark sheet in one of two ways to encourage raters not to focus on single categories (say, content or form) which they may personally favor. The first option would be to overlay the existing hallmark sheet with a number grid (perhaps seven numbers in each category, spanning the three placement levels). This would force raters to evaluate the comparative weaknesses and strengths of the writing and give the raters a visual representation helping them to decide on which side of the placement line a student falls. The other possibility would be to require raters to give at least one comment in each of the five categories on the hallmark sheet. This is conceptually the same as the first option since it forces raters to evaluate each category on the hallmark sheet explicitly. Our suspicion is that raters bring their own priorities and biases into the rating process and at times this may cause them to pay greater attention to certain categories while all but ignoring others. Lumley (2002)

addresses this issue claiming that raters "try to remain close to the scale, but are also heavily influenced by the complex intuitive impression of the text obtained when they first read it. This sets up a tension between the rules and the intuitive impression, which raters resolve by what is ultimately a somewhat indeterminate process." Our recommendation is that the ELI make some minor changes to the hallmark or the qualitative data recording process in order to encourage individual readers to use the hallmark sheet more effectively, thereby helping them to avoid depending too much on their own intuition.

Recommendation #4

This recommendation is incidental to discussions of inter-rater reliability, but we thought it worth mentioning anyway. Like the GRE and TOEFL, we believe that the practices related to the ELIPT should be fully disclosed. This would mean that the Hallmark sheet would be made available to students who are planning to take the ELI placement exam. Sample writing passages that are exemplars of ELI 73, ELI 83, and exempt levels could also be included. Optimally, a rationale for why each of those papers has been placed into their respective courses could also accompany the writing samples.

Recommendation #5

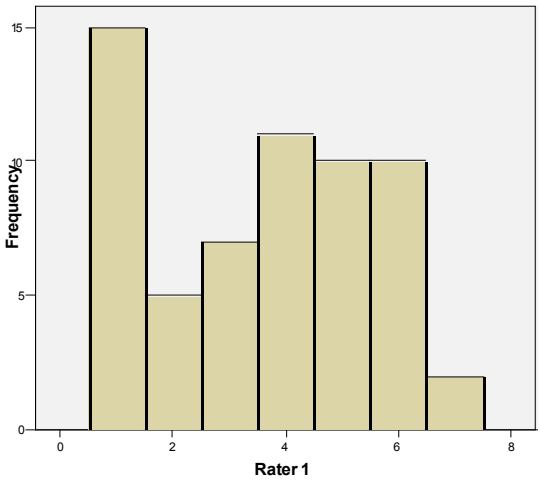
We believe that the training of raters could be somewhat more extensive. We would recommend including the following elements in all future training sessions: 1.) A thorough discussion of coding and oral reporting guidelines. 2.) A discussion of how to use the hallmark sheet and how to report qualitative data. 3.) A guideline for how to prioritize the five categories on the rating sheet. Are all five categories exactly equally important in all circumstances? 4.) Guidelines regarding what to do with grammar errors or other issues which may not be covered by our course objectives. For example, should the presence of article errors in an essay affect

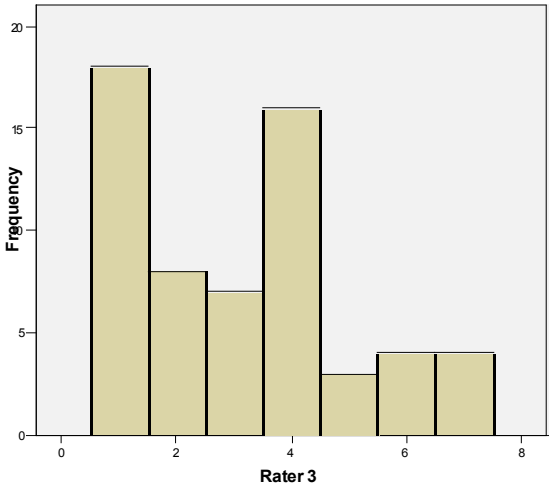
placement? (We normally do not address these errors in our courses, so should we place students based on these errors?)

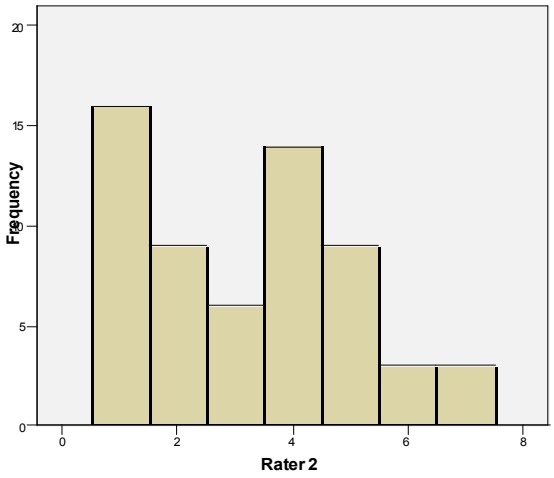
Conclusion

We hope that this project may shed some light on ELI writing placement practices. We feel that we have really only touched the tip of the iceberg and there are many additional analyses that could be performed should future researchers wish to devote their time to this. Inter-rater reliability on the ELI writing placement test was in same ways better and worse than we had expected. Cronbach's alpha reliability estimates at .80 were fairly good, but the size of the standard error of measurement concerned us some. We would be interested in seeing future reliability checks performed and we hope that despite its limitations in staffing and faculty, the ELI will be able to continue to improve this important component of the program.

Appendix A







Appendix B

ELI Writing Hallmarks Version 1.6 21 November 2002

Graduate Students

Level	Content	Organization	Vocabulary	Grammar	Fluency
Exempt Shows high proficiency in L2, knowledge and control of academic writing genres/ conventions common in US universities.	Paper shows evidence of: • Clear point/argument • Complexity of thought/ analysis • Insight on the topic, rather than mere description • Ample supporting evidence, detail, and examples • Command of rhetoric, evidence of writing style	Paper is: • Cohesive • Well-developed • Unique (not formulaic) • Marked by clear transitions, appropriate use of transitional phrases	Paper has: • Wide variety of vocabulary • Appropriate use of idioms • Few problems with collocations • Few problems with word choice	Paper has: • Some errors, but none that interfere with comprehension • Complex sentence structure (e.g., complex coordination, subordination, embedded questions, etc.)	Amount of writing is: • Suitable for level of analysis and/or amount of time provided to write the paper
ELI 83 Shows some knowledge and control of academic writing; needs to develop L2 proficiency, writing ability, and/or awareness of genres/conventions common in US universities.	Paper shows evidence of: • Clear, developed argument, but may be simplistic • Some insight on the topic, but may lack depth • Effective support, but evidence and examples may be general or vague	Paper is: • Cohesive • Somewhat formulaic (e.g. 5-para essay format) • Marked by appropriate transitions, with some misuse/overuse of transitional phrases	Paper has: • Varied vocabulary • Some problems with collocations • Some problems with word choice	Paper has: • Several errors (e.g., tense/aspect, word form, articles, pre-positions), but typically do not interfere with comprehension • Some correct complex sentence structure; evidence of other (incorrect) attempts	Amount of writing is: • Suitable for level of analysis and/or amount of time provided to write the paper
ELI 73 Needs to develop L2 proficiency; notable unfamiliarity with and general lack of control of academic writing; would benefit from at least two semesters of ELI writing instruction.	Paper shows evidence of: • Undeveloped or unclear argument • Simple topic description/ restatement, but with little insight • A general lack of supporting evidence, detail, examples • Redundancy of ideas, argumentation	Paper is: • Not cohesive • Formulaic (e.g., 5-para essay format), or lacking organization • Marked by absence of clear transitions between ideas, or simple sentence-level transitions used at paragraph level (e.g., first, next, then)	Paper has: • Notably limited vocabulary • Repetition/overuse of certain lexical items • Numerous problems with word choice • Incorrect collocations	Paper has: • Numerous errors that typically interfere with comprehension • General lack of sentence complexity	Amount of writing is: • Unsuitable for level of analysis and/or amount of time provided to write paper

Bibliography

Bauchman, L.F., & Palmer, A., S. (1996). *Language Testing in Practice*. Oxford, UK: Oxford University Press.

Crusan, D. (2002). An Assessment of ESL Writing Placement Exam. *Assessing Writing, 8*, pp: 17-30.

Elbow, P. (1993). Ranking, evaluating, and linking: Sorting out three forms of judgment. *College English, 55,* 187-206.

Hamp-Lyons, L. (2001). Fourth generation writing assessment. In: T. Silva & P.K. Matsuda (Eds.), *On Second Language Writing* (pp. 117-127). Mahwah, NJ: Lawrence Erlbaum.

Hamp-Lyons, L. (1997). Exploring bias in essay tests. In: C.Severino, J.C. Guerra, &J.E. Butler (Eds.), *Writing in Multicultural Settings*.New York: The Modern Language Association of America.

Hamp-Lyons, L. (1991a). Issues and directions in assessing second language writing instruction in academic contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 323-329). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1991b). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 323-329). Norwood, NJ: Ablex.

Janopoulous, M. (1992). University faculty tolerance of NS and NNS writing errors: A comparison. *Journal of Second Language Writing, 1,* 109-121.

Johns, A.M. (1997). *Text, Role, and Context*. New York: Cambridge University Press.

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Educaiton, 23*(2), 157-172.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters?. *Language Testing*. *19,*(3), 246-276.

Shor, I. (1997). Our apartheid: Writing instruction and inequality. *Journal of Basic Writing, 16,* 91-104.

Spolsky, B. (1997) The ethics of gatekeeping tests: What have we learned in a hundred years?: *Language Testing, 14,* 242-247.

Sweendler-Brown, C.O. (1993). ESL essay evaluation: The influence of sentence level and rhetorical features. *Journal of Second Language Writing, 2,* 3-17.

Vann, R.J., Lorenz, F.O., & Meyer, D.M. (1991). Error gravity: Faculty response to errors in the written discourse of nonnative speakers of English. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 181-195). Norwood, NJ: Ablex.

Vann, R.J., Meyer, D.M., & Lorenz, F.O. (1984). Error gravity: A study on faculty opion on ESL errors. *TESOL Quarterly, 18*, 427-440.

White, E., M. (1994). *Teaching and Assessing Writing* (2nd ed.).San Francisco: Josey-Bass.