Fall 10-21-2011

Exploring Language as a Source of DIF in a Math Test for English Language Learners

Minji K. Lee
University of Massachusetts Amherst, minjik@educ.umass.edu

Jennifer Randall
University of Massachusetts Amherst, jrandall@educ.umass.edu

Running head: LANGUAGE AS A SOURCE OF DIF

Exploring Language as a Source of DIF in a Math Test for English Language Learners

Minji Lee

Jennifer Randall

University of Massachusetts, Amherst

Exploring language as a source of DIF in a math test for English language learners

Abstract

English language learners (ELs) have shown lower performance in mathematics than non-ELs although mathematics is an area that uses the least amount of language among the subjects that are mainly tested. If this differential performance is due to the bias in test items, then validity of using ELs' test scores in comparison to non-ELs' is compromised. For this reason, studies have investigated whether the differential performance can be attributed to language load in the tests. The results of these studies were not consistent. Some studies did find its effect, whereas others did not. Some of the difficulties encountered by researchers in past studies investigating DIF include a large difference in sample size between the two groups and unclear distinctions between ELs and non-ELs. This study aims to investigate the source of DIF between ELs and non-ELs using a comparatively large and a better defined/restricted population of ELs. This study will contribute to existing knowledge about English proficiency as a possible cause of differential performance between the two groups. The findings of this study will have implications for test construction and policies for providing testing accommodations (e.g., test language simplification)

Introduction

English language learners (ELs) score lower on average than non-ELs on math tests. Many studies have identified language as a source of differential performance between ELs and

non-ELs (Abedi, 2002; Abedi et al. 2005; Abedi & Gandara, 2006; Abedi, Lord, and Plummer, 2006, Barton & Neville-Barton, 2003; Eid, 2002; Martiniello, 2009; Pomplum & Omar, 2001; Wolf & Leon, 2009). A problem raised by this finding is that if the differential performance in math is attributable to difficulty in understanding the written language in math tests, rather than to a difference in math ability, then we cannot make valid inferences about ELs' math ability from their math test scores (AERA, APA & NCME, 1999). For example, if certain math problems contain technical vocabulary and complex syntax, these problems may be harder for ELs than for non-ELs with equivalent math ability. For this reason, the extent to which a given math test presents construct-irrelevant variance associated with English language proficiency has been of great interest to researchers (Abedi, 2002; Abedi et al., 2005; Abedi, Hofstetter, Baker, & Lord, 2001; Abedi & Lord, 2001; Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, & Plummer, 1997; Eid, 2002; Johnson & Monroe, 2004; Mahoney, 2008; Martiniello, 2008, 2009; Ockey, 2007; Shaftel et al. 2006; Walker, Zhang, & Surber, 2008; Wheeler & McNutt, 1983; Wolf & Leon, 2009).

All cases of differential math performance, however, may not be a function of inherent bias against ELs (Mahoney, 2008; Ockey, 2007). Inherent bias refers to *differential item functioning* (DIF), which is a psychometric characteristic of an item that can misrepresent the competence of one group (Shepard, 1982) as explained in the previous paragraph. Whereas DIF refers to differences in item functioning *after* groups have been matched with respect to the ability, *impact* reflects differences in overall ability distributions between two intact groups (Dorans & Holland, 1993). For example, seniors usually score higher than junior high school students on typical SAT mathematics items.

In the context of this study, the score gap in math between ELs and non-ELs may simply be a manifestation of impact rather than DIF. For example, ELs may perform worse than non-ELs in mathematics assessments not because of the linguistic complexity of the test, but because they often have differential exposure to the math curriculum, for example, as a result of their prolonged placement in ESL classrooms. In fact, a number of researchers have discussed the relationship between the reduced opportunity to learn for ELs and their low performance on math assessments (Abedi and Herman, 2010; Abedi and Gandara, 2006; Abedi, Lord, and Plummer, 1997; Callahan, Wilkinson, and Muller, 2010). If ELs learned the same math content as non-ELs, the score gap between these ELs and non-ELs may possibly be minimal.

However, this does not rule out the hypothesis that language factors compromise the interpretation of ELs' math test scores. Therefore, it would still be important to provide evidence that the performance gap between ELs and non-ELs does not reflect test bias, and one way of doing so would be to show that there is no substantial DIF between ELs and non-ELs due to language factors. In order to confidently attribute the differential performance between ELs and non-ELs to language barriers, one must investigate whether any items show substantial DIF and also whether the linguistic complexity of an item predicts the magnitude of any DIF that is found.

Some of the difficulties encountered by researchers in past studies investigating DIF include a large difference in sample size between the two groups (e.g., 34 ELs against 1,060 non-ELs), and unclear distinctions (i.e., overlapping English proficiency distribution) between ELs and non-ELs (Johnson & Monroe, 2004; Ockey, 2007; Shaftel et al. 2006). This study will address these problems by investigating the source of DIF between ELs and non-ELs using a comparatively large and a better defined/restricted population of ELs. The EL group will be

divided into two subcategories according to their English language proficiency (e.g., limited English proficient (LEP) and former LEPs) and comparisons will be made for each of the groups.

Goal and Research Questions

The purpose of this study is to compare the performance of ELs in a statewide math test to that of non-ELs, to determine if linguistic complexity is a source of DIF between ELs and non-ELs. This study will contribute to existing knowledge about English proficiency as a possible cause of differential performance between the two groups. The relevant research questions in exploring this problem are: (1) Does the linguistic complexity of an item predict the magnitude of DIF in a math test? (2) Does comparing former LEPs with non-ELs lead to fewer instances of DIF in a math test? (3) Does comparing former LEPs with non-ELs lead to smaller DIF effects?

Background

Eight studies were identified, which focused on test item language as a source of DIF between ELs and non-ELs; four studies identified language complexity in the test as a source of DIF (Eid, 2002; Martiniello, 2008, 2009; Wolf & Leon, 2009), whereas the other four failed to find a relationship between language complexity and DIF (Mahoney, 2008; Miller, Doolittle, & Ackerman, 1988; Ockey, 2007; Snetzler & Qualls, 2000). These last eight studies are introduced in the following paragraphs and the limitations of these studies will be discussed.

Eid (2002) investigated item characteristics that might produce DIF in the mathematics part of the SAT between students who speak English as their best language (EBL) and non-EBL students. The results of the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988) for EBL and non-EBL groups identified 12 “C” (i.e., large-DIF) items with four items favoring non-EBLs and eight items favoring EBLs, and 12 “B” (i.e., moderate-DIF) items with seven items favoring

non-EBLs and five items favoring EBLs. He found significant relationships between item difficulty and the DIF statistic, α_{MH} (r = .29, p = .02), and between item readability grade level and α_{MH} (r = -.86, p < .001). In addition, readability grade level accounted for 80% of the variability in α_{MH}, which was a statistically significant amount of variability in the dependent variable, α_{MH}. In addition, the α_{MH} mean was higher for low readability items, and items with high readability tended to favor the focal group, whereas items with low readability tended to favor the reference group.

Martiniello (2008) described linguistic features of fourth-grade math items from the Spring 2003 Massachusetts Comprehensive Assessment System (MCAS) that showed DIF against ELs. Among the 39 items, nine were identified as B-DIF and one as C-DIF. Only two of these ten items exhibited meaningfully large DIF. The author examined six of the ten items that favored non-ELs over ELs in order to assess the possible contribution of linguistic complexity to the difficulty experienced by ELs.

Using textual analysis and children's think-aloud transcripts, the author identified linguistic features of DIF items that disfavored ELs. Syntactic features of this kind included multiple clauses, long noun phrases, and a lack of clear relationships between the syntactic units. Lexical features included sophisticated academic words, words usually learned at home, words with multiple meanings, expressions that portray particular aspects of mainstream American culture, and a lack of correspondence between the syntactic boundaries of clauses and the layout of the text. She stated that the empirical evidence tends to confirm the hypothesis that linguistic complexity is a source of DIF.

Martiniello (2009) also investigated the relationships among linguistic complexity, schematic representation (i.e., symbols and figures), and DIF in the same sample described

above. The results indicated that the effect of linguistic complexity on DIF disfavoring ELs was significant and positive ($p < .001$). There was also a significant interaction between linguistic complexity and the presence of schematic representations in the item ($p < .015$). That is, the overall impact of linguistic complexity on DIF was attenuated when ELs could rely on nonlinguistic schematic representations. The author suggested that the inclusion of schematic representation could help mitigate the negative effect of increased linguistic complexity on EL math performance.

Wolf and Leon (2009) also found linguistic complexity as a key variable in explaining DIF. Academic vocabulary was a prominent feature characterizing linguistic complexity. General academic vocabulary was more likely to cause DIF than other types of vocabulary among items requiring relatively easy content knowledge. The authors suggested that this finding is consistent with ELs having explicit opportunities to learn technical and context-specific vocabulary, while lacking opportunity to learn general academic words such as *based on* or *substantial* during content instruction. Additionally, more DIF was present when the focal group consisted of ELs with low English proficiency rather than high proficiency, and for easy items than for hard items. For relatively easy items, higher linguistic complexity was associated with greater uniform DIF against ELs. For hard items, the pattern was inconsistent, which indicated that factors other than linguistic complexity might influence DIF.

The authors suggested that these findings have important implications for test validity and test development as well as for instruction for ELs. A test with a more number of easy items tended to exhibit more DIF items, implying that the linguistic complexity was more likely to lead to DIF for ELs in easy items. However, including easy items may be necessary to better discriminate among low-performing students such as ELs. In addition, this study has an

implication for the use of linguistic simplification. Linguistic simplification may reduce linguistic obstacles for EL students taking a content test, but may also remove context-specific and technical vocabulary items, which is a part of the construct being tested. This may lead to construct underrepresentation, which raises questions about the validity of the score-based interpretations of the math test. According to their findings, general academic vocabulary tended to cause DIF rather than technical academic vocabulary, so the authors suggested that test developers should be mindful of using general academic vocabulary, which is typically not a part of the construct to be measured. The different DIF findings between the high- and low- EL samples confirmed that EL population is a highly heterogeneous group, which has implications for policies for providing testing accommodations. Finally, the authors recommended refining and applying the linguistic coding scheme and the DIF methods utilized in this study for validating assessments for EL students.

In contrast to the previous four studies summarized, Ockey (2007) could not confirm the hypothesis that language ability is a cause of DIF with respect to math word problems. In his study, the difference in test performance between ELs and non-ELs was statistically significant ($t = 9.05$, $p < .001$) and practically important (Cohen's $d = .57$). However, the existence of an extraneous second trait, such as language ability, was not supported by the principal components analysis. He also found only one item displayed against ELs using the IRT procedure and MH DIF analysis. Given this finding, the author suggested that math word problems do not inherently contain DIF against ELs. The author also stated that carefully crafted math word problems can be fairly used to assess the abilities of ELs.

However, Ockey recognized the possibility that statistical techniques used in this study might not have been effective in identifying DIF against ELs. He did not delineate exactly what

aspects of the statistical techniques were ineffective in this study, but he did state that a sharp distinction between ELs and non-ELs was absent. In addition, he pointed out the small sample size of the focal group, and the small number of items. He recommended using different DIF designs in the future, by making the EL/non-EL distinction more salient by comparing native or near-native English speakers with low-level ELs, or by using an independent measure of math ability as the conditioning variable rather than the ability measured by the items on the test one is investigating.

Mahoney (2008) used confirmatory factor analysis to examine the effects of potential irrelevant constructs on math achievement within the context of the 1996 NAEP mathematics assessment administered to fourth graders. The goal of this study was to investigate items that display DIF between second language learners (SLL) and English-only (EO) students, and the relationship between items' linguistic complexity and DIF. The author used multiple-group CFA to compare latent-variable models for the 25 items. A single-factor model was fit, because all 25 items were hypothesized to measure math achievement. A baseline model with no constraints between the two groups, a constrained model with invariant loadings, and a model with invariant thresholds in addition to invariant loadings were each fit to data. It was expected that goodness-of-fit indices would decline as constraint levels increased. The presence of DIF is indicated by a significant and meaningful decline in the fit indices.

The author confirmed invariance in factor loadings and thresholds between SLL and EO students, and concluded that items functioned similarly for both groups despite the variation in the linguistic complexity of the items. However, she noted large standard errors and conjectured that the sample may not have been large enough for the methodology employed. Second, she mentioned some cases of misfitting items found in the baseline model as a possible cause of

undetected DIF. Lastly, she pointed out an important limitation of her study: the unknown English language-proficiency levels of the SLLs. It is possible the distributions of EL proficiency in the SLL and EO groups largely overlapped. This illustrates why a large group of well-defined ELs is necessary.

Miller, Doolittle, and Ackerman (1988) investigated whether mathematics items with the greatest verbal load tend to favor non-ESL examinees. They examined 40 items that measure mathematical reasoning ability in six content areas from the American College Testing Program Assessment (ACT). In general, the hypothesis that high word-count items favor non-ESL students was not supported. However, the authors noted that the two DIF items that favored non-ESL students were high word-count items. However, the authors judged that the evidence was not conclusive, because a constant group difference in math performance was reflected throughout most of the items in the test. Specific categories of items that were disproportionately easy or difficult for either group could not be found. A different conclusion may have been reached if the focal group had been larger in size.

Using the Iowa Tests of Basic Skills (ITBS), Snetzler and Qualls (2000) examined the possible presence of DIF between ELs and non-ELs using MH procedure. First, the authors compared the limited English to bilingual students. The tests as a whole appeared to be quite difficult for both groups. Among the fourth-graders, the effect size of the performance difference between groups was more than half a pooled standard deviation favoring bilinguals. For this same group of students two years later, the effect size increased to 0.74. Among the sixth-graders, on the other hand, there was no substantial performance difference at either time point. The authors conjectured that the inconsistencies were due to an inappropriate level of difficulty. There were no C-DIF items in any test. For B-DIF items, there were fluctuations as to which

group was favored. The authors attributed the inconsistencies found in this study to the extreme difficulty that these tests presented for examinees.

The last four studies did not find consistent patterns of DIF against ELs and thus could not find evidence of language as a source of DIF. The commonality shared by these studies seems to be the small number of ELs. Ockey (2007), Miller, Doolittle, and Ackerman (1988), and Snetzler and Qualls (2000) had focal groups fewer than 500, and Mahoney (2008) had fewer than 1000. In addition, two studies that employed IRT and SEM for detecting DIF did not address possible differences in proficiency distributions between the groups. However, if sample size is inadequate, and if there are differences between groups in the distribution of proficiency, we cannot be sure whether any findings with respect to measurement invariance, positive or negative, are the result of genuine qualitative differences. In addition, some studies pointed to the possible failure to make crisp classifications of ELs and non-ELs. The present study will address these issues by acquiring a larger sample of ELs, employing a stratified random sampling technique to account for differing proficiency distributions, and making comparisons using LEP and former LEP groups distinguished by their English ability.

Method

Participants

The population consists of students in an eastern state who sat for the state achievement examination in mathematics. Matched sampling technique will be used to match examinees from the reference group (non-EL) to examinees from the focal group (EL) according to mathematics ability. The proxy for ability will be the sum score on the mathematics test. For each EL with a given sum score, *n* non-ELs who have the same sum score will be randomly drawn from the

population. This sampling method will result in equivalent distributions of sum scores across the two groups, which helps to ensure that mathematics ability is similarly scaled regardless of group. Sampling more than one non-EL student for each EL student is expected to provide more stable parameter estimates. This sampling technique will be used for DIF analysis using NOHARM. With the MH procedure, a previous study found that comparing all examinees from the reference group to the focal group is a preferred strategy across different distributional conditions (Lee, Wells, & Sireci, 2011). The ELs will be divided into two subgroups according to their English proficiency (LEPs and former LEPs) for more refined DIF analyses. Two comparisons will be made in total: (1) LEPs vs. non-ELs, and (2) former LEPs versus non-ELs.

Instrument

The math scores were obtained from the results of the statewide achievement test. The measure of the linguistic complexity of the mathematics items came from a linguistic analysis protocol created for this study. Four raters were trained to use the protocol, which contains scoring rubrics for lexical and grammatical components. The lexical rubric asks the raters to count the frequency of general academic vocabulary (e.g., *consequently*, *based on*) and judge the impact of low-frequency words. Examples of general academic vocabulary were provided to the raters. The rationale for including general academic vocabulary as a component of lexical analysis is based on Wolf and Leon (2009)'s finding of an association between general academic vocabulary and DIF. Mathematics vocabulary was excluded from the linguistic lexical analysis because math vocabulary is a part of the construct measured by the test. Two mathematics teachers judged whether certain vocabulary is content-relevant. The raters were also asked to consider whether the meaning of the low-frequency words are hard to derive from the context, and whether ignorance of a word's meaning would interfere with comprehension of the problem.

On the basis of the two categories (i.e., the frequency of general academic vocabulary, the impact of low-frequency nonmathematical words), the raters will be asked to score the lexical complexity for each item on a Likert-type scale ranging from 1 (least complex) to 5 (most complex).

Next, the grammatical rubric involved counting features such as passive voice phrases, nominalizations, modals, conditional and relative clauses, and total number of words, sentences, and words per sentence. The first five features are commonly recognized as features of academic English (Butler et al., 2004; Schleppegrell, 2001; Wolf and Leon, 2009). The raters were also asked to score the grammatical complexity of each item on a Likert-type scale ranging from 1 (least complex) to 5 (most complex). The ratings of each item were averaged across the five raters to produce a mean item lexical complexity score and a mean item grammatical complexity score. Mathematics items from a test other than the one used for this study were analyzed first for training purposes. Interrater agreement will be computed as a check on the reliability of these two measurements.

DIF Analysis

Because there are numerous approaches to DIF detection, each relying on different statistical techniques, it is prudent to use more than one method to ensure that results are robust. Two kinds of DIF analyses were conducted: (1) McDonald's modification of Lord's procedure, which makes use of nonlinear factor analysis (Lord, 1980; McDonald, 1999), and (2) MH procedure.

DIF detection using NOHARM procedure. The computer program NOHARM conducts nonlinear factor analysis of dichotomous items (Fraser & McDonald, 2003). The output of NOHARM includes each item's factor loading (λ_j) and threshold (τ_j). In the model employed

by NOHARM, each item score is posited to be a dichotomization of an underlying continuous quantity. If an examinee's amount of this quantity exceeds the item's threshold, the examinee gives the correct response to the item; if not, then the examinee gives the incorrect response. The factor loadings given by NOHARM are in fact the loadings of the continuous quantities underlying each item on the common factor measured by the test as a whole. NOHARM places the sign on the threshold parameters in such a way that larger values correspond to easier items; the area under the normal curve from negative infinity to the threshold parameter gives the probability of an examinee responding to an item correctly. Thus, each τ_j can be treated as a Z-score of item difficulty, which helps with statistical testing. McDonald (1999) showed that this nonlinear factor analysis is mathematically equivalent to IRT. In IRT, the probability of an examinee with factor level θ getting item j correct is equal to

$$P(x_j = 1 | \theta) = \frac{1}{1 + \exp[-1.7 a_j (\theta - b_j)]},$$

where a_j is the discrimination parameter and b_j is the difficulty parameter. McDonald showed that λ_j can be converted to a_j using the equation $a_j = \lambda_j / \sqrt{1 - \lambda_j^2}$. We can use the equation $b_j = -\tau_j / \lambda_j$ to obtain b_j.

Lord (1980) initially proposed plotting one group's values of an IRT parameter against the other group's values. Any items lying far from the best-fitting straight line are identified as showing DIF. McDonald modified this procedure by using the factor-analytic parameterization of IRT. McDonald found in his example that the factor-analytic parameterization tended to agree more across groups than the traditional parameterization, which suggested that the former produces more accurate estimates. This led McDonald to conjecture that the factor-analytic parameterization is superior for the purpose of conducting DIF analysis.

In the context of NOHARM, DIF can occur for two reasons: (1) the two groups have different thresholds (τ_j), or (2) the two groups have different factor loadings (λ_j). Both a visual and statistical hypothesis-testing approach were used to look for DIF. First, in the visual approach, if there is no DIF and the groups have identical ability distributions, then in a graph plotting the factor loadings (or thresholds) in one group against the factor loadings (or thresholds) in the other group, the points should lie on a straight line through the origin with a slope of one. If the item parameters do not show this pattern as a whole, then we can conclude that there are many items showing DIF. This visual approach was used to identify DIF items.

A statistical hypothesis test was also used to look for DIF. The standard errors for factor loadings and thresholds were computed to test the differences in the parameters between the two groups. According to McDonald, the standard error of a factor loading estimated by NOHARM is approximated by the reciprocal of the root sample size. The matched sampling used in this study makes the sample sizes of the groups being compared equal, which makes the standard errors for the factor loadings in the two groups equal. The standard error of the difference between the two factor loadings can be computed as

$$\mathrm{SE}(\hat{\lambda}_R - \hat{\lambda}_F) = \sqrt{Var(\hat{\lambda}_R - \hat{\lambda}_F)}$$

$$= \sqrt{Var(\hat{\lambda}_R) + Var(\hat{\lambda}_F) - 2Cov(\hat{\lambda}_R - \hat{\lambda}_F)}$$

$$= \sqrt{Var(\hat{\lambda}_R) + Var(\hat{\lambda}_F)} = \sqrt{(1/N_R) + (1/N_F)}\,.$$

where $\hat{\lambda}_R$ stands for the estimate of the factor loading for the reference group and $\hat{\lambda}_F$ is the estimate of the factor loading for the focal group. This procedure can be used to construct the 95% confidence interval $\hat{\lambda}_R - \hat{\lambda}_F \pm 1.96 \times \sqrt{2} \times SE(\hat{\lambda}_R)$ to for the discrepancy in factor loadings.

The difference in threshold parameters can be easily transformed to the proportion getting the item correct. Therefore, confidence interval for the differences between two proportions will be used to test for group discrepancies in thresholds. This procedure involves computing the standard error of the difference between two proportions, $SE = \sqrt{\frac{\pi_r(1-\pi_r)}{n_r} + \frac{\pi_f(1-\pi_f)}{n_f}}$, where π_r and π_f are the proportions of the reference group members and the focal group members answering an item correctly, n_r is the size of the reference group, and n_f is the size of the focal group.

To determine whether the differences are large enough to be a cause for concern, the graphical displays of the differences between trace lines were examined (Steinberg and Thissen, 2006). Therefore, the test characteristic curves (TCCs) of the student groups were obtained for each of the five content strands to gauge the effect size of the differences.

DIF Detection via the Mantel-Haenszel Method. The Mantel-Haenszel (MH) method for detecting DIF was implemented with the software package R (R Core Development Team, 2010). The MH method tests whether the odds of getting an item correct at a given level of the matching variable is the same in both the focal group and the reference group across all levels of the matching variable. The MH method is based directly on observable statistics and thus is a non-parametric technique in that it does not require estimation of model parameters. An odds ratio of one indicates null DIF; an odds ratio different from one indicates DIF. The pooled estimate of the odds ratio across levels of the matching variable, α_{MH}, is an estimate of DIF effect size on a metric that ranges from 0 to ∞ with a value of one indicating null DIF. α_{MH} is usually converted to log odds because the latter is symmetric around zero and easier to interpret. Since Educational Testing Service (ETS) uses the *delta metric* for item difficulties, which is

normal with a mean of 13 and a standard deviation of 4, Holland and Thayer (1985) converted α_{MH} into a difference in deltas via $MH\ D\text{-}DIF = -2.35\ln[\alpha_{MH}]$. This gives a suitable estimate of an effect size. A classification scheme developed by ETS, categorizing items as exhibiting negligible DIF (A-DIF), intermediate DIF (B-DIF), or large DIF (C-DIF), was used in this study. Items that exhibited B-DIF or C-DIF were flagged. Specifically, if the *p*-value for the null hypothesis of no DIF was greater than .05, or if the *p*-value was less than .05 but the absolute delta effect size was less than 1.0, then the item was categorized as no or negligible DIF (A-DIF). An item was flagged as B-DIF if the *p*-value was less than .05 and the absolute delta effect size was between 1.0 and 1.5. An item was flagged as C-DIF if it showed a *p*-value less than .05 and an absolute delta effect size greater than 1.5. The MH method can only test for an overall odds-ratio difference across ability levels. Therefore, one disadvantage of the MH method is that it can only detect uniform DIF. The purpose of conducting MH DIF analysis was to cross-validate the results obtained from NOHARM analysis, especially for uniform DIF.

The items showing DIF using both the NOHARM and the MH methods were identified and examined with respect to their linguistic complexity. If linguistic complexity turns out to be a significant predictor of discrete DIF status or continuous measures of DIF effect size (in linear regression), then our study will shed light on our research questions.

Impact

(1) This study will contribute to existing knowledge about English proficiency as a possible cause of differential performance between the two groups.

(2) The findings of this study will have implications for test construction. Item writers can be mindful of the language variables if they are found to be significant sources of DIF.

Results

Descriptive Statistics

Table 1 shows the descriptive statistics of the sum scores that non-EL, LEP, and former LEP (FLEP) students obtained on the math test. The sample sizes of the groups were 60,000+, 2,844, and 1,314 respectively. As expected, the non-EL students had the highest mean score (M = 22.56), and the FLEP group (M = 17.9) scored 3.9 points higher on average than the LEPs (M = 14). Figure 1 shows the score distributions of the three groups. The score distribution of the non-ELs was negatively skewed, whereas the distributions of the LEPs and former LEPs were positively skewed.

Matched Sampling

As pointed out in the earlier section, if two groups under comparison differ greatly in their ability distributions, matched sampling conditional on total scores may help ensure that any findings of DIF are not due to the distributional differences. Therefore, matched samples of the non-EL population were used for the NOHARM analysis. Two matched samples were created, one matched to the score distribution of the LEPs and the other to that of the FLEP students. Note that we sampled *n* reference group members for each focal group member. For non-ELs matched to LEPs, *n* was 2. For non-ELs matched to FLEP, *n* was 15. In both cases, the matched non-EL samples came to have the same means, variances, and distributional shapes of sum scores as the focal groups.

NOHARM Data Analysis

A two-dimensional two-parameter IRT model for the LEP population was estimated to test the hypothesis that there is an extra language dimension in the math test. The Tanaka goodness-of-fit index was .992, indicating a good fit. However, a unidimensional model fit as

well as the two-dimensional model, with a Tanaka GFI of .992. This suggests that additional dimensions had little influence on examinee responses. A unidimensional model fit well for the FLEP population with a GFI of .992. Similarly, a unidimensional model fit well for the non-EL samples matched to the LEP and the FLEP population, showing GFI values of .992 and .995 respectively.

Next, in order to test DIF, we computed the differences in thresholds and factor loadings, the distributions of which are presented in Figure 2. The magnitudes of the differences are smaller in the non-EL vs. FLEP comparison than in the non-EL vs. LEP comparison. There were some observations that lay slightly far from the overall distributions, such as items 11 and 12 for thresholds and items 17, 29, and 21 for factor loadings. These items would be identified as potential DIF items in the later analyses.

DIF with respect to thresholds. First, we conducted a statistical test for the significance of the differences in thresholds. With large enough sample sizes, significance tests will always find that differences in the estimates are statistically significant. With sample sizes as large as those used in this study (see Table 1), many of the differences in factor loadings and thresholds could be identified as statistically significant. Indeed, in the LEP vs. non-EL comparison, as many as 31 out of 34 items were identified as having significantly different thresholds. In the FLEP vs. non-EL comparison, 27 items were identified as having significantly different thresholds.

Now we take a graphical approach. If there is no DIF, then in a graph plotting the estimated parameters of one group against those of another, the points should lie on a straight line through the origin with a slope of one. Figure 3 shows such a graph for thresholds in the LEP and non-EL samples. Figure 4 shows the corresponding graph for the FLEP and non-EL samples.

The fit to a straight line through the origin with a slope of 1 looks fairly good in both plots. In the first plot, a linear regression of the LEP group's thresholds on the non-EL group's thresholds gave an intercept of -0.01 and a slope of 0.98. In the second plot, a linear regression of the FLEP group's thresholds on the non-EL group's thresholds gave an intercept of 0 and a slope of 0.99. This good fit to a straight line through the origin with a slope of 1 means that there is not much overall DIF at the test level with respect to the thresholds (item difficulties). There were some signs of potentially noteworthy DIF with respect to thresholds in items 11, 12, 20, 24, 31, and 34 in the comparison of the LEP and non-EL groups, and in items 11 and 12 in the comparison of the FLEP and non-EL groups. In both comparisons, item 11 was easier for the focal group. The other items (12, 20, 24, 31, and 34) were easier for the reference group.

DIF with respect to factor loadings. According to a statistical test for the significance of the differences in factor loadings, as many as 19 items were identified as having factor loadings significantly different from each other in the LEP vs. non-EL comparison. There were fewer items—only five—identified as having significant differences in factor loadings in the FLEP vs. the non-EL comparison.

In the graphical approach, the fit to a straight line through the origin with a slope of 1 again looks fairly good in both Figure 5 and Figure 6. In fact, a linear regression of the LEP group's factor loadings on the non-EL group's factor loading gave an intercept of -0.03 and a slope of 1.06; and a linear regression of the FLEP group's factor loadings on the non-EL group's factor loadings gave an intercept of -0.07 and a slope of 1.12. This good fit to a straight line through the origin with a slope of 1 means that there is not much overall DIF at the test level with respect to the factor loadings. However, there were some signs of potentially noteworthy DIF with respect to factor loading in items 3, 18, 22, 29, 31, 32, and 34 in the LEP vs. non-EL

comparison, and in items 17 and 29 in the FLEP vs. non-EL comparison. In both comparisons, items 29, 31, 32, and 34 were less discriminating for the focal groups, whereas items 17, 21, and 22 were less discriminating for the reference group.

MH DIF procedure

Uniform DIF. As stated previously, the MH DIF procedure allows us to test for differences in difficulty, but not in discrimination. In the non-EL vs. LEP comparison, items 11, 12, 20 and 31 were identified as showing meaningful DIF. Among them, item 12 was a C-DIF (i.e., severe DIF) item. In the non-EL vs. FLEP comparison, item 12 was again identified, but the magnitude of the DIF effect size became smaller, showing B-DIF in this comparison.

DIF analysis using a purified criterion. The sum of all dichotomous items, including DIF items, can be an inadequate matching criterion. Therefore, additional MH analyses were conducted using a purified criterion without the DIF items. Exactly the same items were identified as showing a given level of DIF as in the MH analyses using a non-purified criterion.

Overall results of DIF analyses

Table 3 summarizes the overall results of our DIF analyses. Eleven items were identified as showing potential DIF with respect to either thresholds or factor loadings. In the non-EL vs. LEP comparison, the NOHARM nonlinear factor analysis identified six items showing DIF with respect to thresholds. Four of these were also identified by the MH analysis. In the non-EL vs. FLEP comparison, only a subset of the DIF items from the previous analyses was identified; the NOHARM approach flagged two items, and the same items were also flagged by the MH analysis. With respect to factor loadings, seven items were identified by the NOHARM analysis in the non-EL vs. LEP comparison. In the non-EL vs. FLEP comparison, fewer items were identified.

In interpreting these DIF results, it is important to consider the content area represented by each item. As many as *five* DIF items came from the content area *Data Analysis, Statistics, and Probability.* Note that the total number of items in this content area was seven. These items were systematically harder or less discriminating for the focal group. One of the items (12) belonging to this content strand showed an extreme level of DIF with respect to threshold. The other six items were either from *Number Sense and Operations* or *Patterns, Relations, and Algebra.*

Test Characteristic Curves

Figure 6 shows the test characteristic curves (TCCs) of the five content areas for non-ELs and LEPs, and Figure 7 shows these curves for non-ELs and FLEPs. The TCCs for the first two content areas are very similar for non-ELs and LEPs. For example, the maximum expected score difference between the groups in *Geometry* was 0.11; in *Measurement,* it was 0.12. There may have been small amounts of DIF in the next two content areas. The maximum expected score difference between groups in *Number Sense and Operations* was 0.41; in *Patterns, Relations and Algebra*, it was 0.35. In both cases, the LEP group was favored. In *Data Analysis, Statistics, and Probability,* however, the LEPs experienced a noticeable disadvantage. The maximum expected score difference between the groups in this content area was 0.77. Could the disadvantage experienced by LEPs in this content area be explained by the linguistic complexity of those items? For example, Martiniello (2008) found that two DIF items concerning probability and statistics had relatively lengthy sentences and unfamiliar vocabulary.

Linguistic complexity of items

Expert ratings. As we can see from Figure 8 and Figure 9, the lexical and grammatical complexities of the math items were low. The maximum rating of five was never used. One and

two were by far the most frequent ratings. Table 4 shows the full descriptive statistics of the rated linguistic characteristics of the items. Mathematical terms (e.g., *correlation*, *scatterplots*, *slope*, *parallel*) were excluded from consideration of linguistic complexity, because they were considered a part of the construct being tested.

Most of the items were rated quite low in terms of lexical complexity. There were several general academic words that students tend to encounter across subjects (e.g., *represent*, *following*, *based on*) and low-frequency expressions (e.g., *velvet*, *wingspan*, *life span*). When low-frequency words did occur, their meanings were often judged easy to derive from context. If they were hard to derive from context, they were mostly judged not to interfere with understanding of the text. Most of the items with low-frequency words that were judged to interfere with understanding of the text did not exhibit DIF. There was one exception: One rater indicated that not knowing the meaning of the word *record* would interfere with comprehension of an item from the *Data Analysis, Statistics, and Probability* content strand. This item asked the probability of selecting an object at random with replacement, each time recording which object was selected. This item did exhibit DIF in its factor loading in favor of non-ELs.

Most of the items were also rated low in terms of grammatical complexity. Passive voice phrases, conditional or relative clauses, modals, and nominalizations occurred infrequently. After rounding the mean number of sentences in an item was 2, and the mean total number of words in an item was 30.

Note that the intraclass correlation of the lexical complexity rating was 0.31 with confidence interval (0.14, 0.52); the intraclass correlation of the grammatical complexity rating was 0.42 with confidence interval (0.24, 0.61). The interrater reliability estimates may have been low partly as a result of small variance in “true scores”; the typical item elicited a rating of one

from almost all raters. Since every item has almost the same true score, which is close to one, the variance of errors becomes more prominent.

Multivariate Analysis of Variance analysis. A MANOVA analysis was conducted to determine whether ratings of lexical and grammatical complexity differed by content area. The effect of content was not statistically significant (Wilks lambda = .83, p =.69). In addition, none of the comparisons (Algebra vs. Statistics, Number Sense vs. Statistics, Geometry vs. Statistics, and Measurement vs. Statistics) was significantly different from zero with respect to either lexical or grammatical complexity. Therefore, the disadvantage experienced by LEPs in the content area *Data Analysis, Statistics, and Probability* could not be explained by linguistic complexity.

Regression models. None of the linguistic predictors explained the differences in either the thresholds or the MH DIF effect sizes in comparison of non-ELs and LEPs (or FLEPs) (Figures 10 and 11). This remained true in multiple regressions with six of the linguistic predictors as shown in Models 1 and 2 in Table 5. In contrast, as shown in the simple linear regressions in Figure 12, there were several linguistic variables that had statistically significant relationships with factor loading DIF in comparison of non-ELs and LEPs. For example, as lexical complexity or grammatical complexity increased to a certain point, items tended to have higher factor loadings (i.e., to be better indicators of math ability) for non-ELs relative to LEPs. As linguistic complexity increased past this point, the difference in the factor loadings declined. There was also a statistically significant linear relationship between the total number of words and the differences in factor loadings between non-ELs and LEPs. As the number of words in an item increased, it tended to become less effective as an indicator of math ability for LEPs relative

to non-ELs. The linguistic complexity variables did not account for the differences in factor loadings in non-EL vs. FLEP comparison.

Models 3 and 4 show the results of multiple regression analyses. Model 4 was fit excluding items 18, 28, and 31, which were far from the straight line in the Q-Q plot of the initial residuals. In this model, grammatical complexity and the total number of words remained significant predictors of differences in factor loadings. As the total number of words or lexical complexity increased, with other predictors fixed to some constants, items became less discriminating for LEPs relative to non-ELs. Interestingly, after statistically controlling for the fixed total number of words and lexical complexity, grammatically more complex items tended to be more discriminating for LEPs than for non-ELs.

Discussion

Linguistic complexity has often been perceived as a potential influence on the math performance of EL students. This study sought to investigate linguistic complexity as a source of DIF by using a relatively large EL group and dividing it to LEP and FLEP by English proficiency. It also addressed a methodological issue by matching two groups in terms of score distribution, thus helping to ensure that the DIF findings were not reflective of a difference in ability distribution between the groups under comparison. In the statewide achievement test investigated in the current study, a large proportion of the items from the content area *Data Analysis, Statistics, and Probability* showed DIF against LEPs. The number of DIF items and the magnitude of DIF became smaller for FLEPs, and this content strand did not appear to put this group at a serious disadvantage. MANOVA analyses indicated that the items from this content strand were not more linguistically complex than items from other content strands. A possible explanation for DIF against LEPs in this content strand may be differential educational exposure.

If LEPs went through the math curriculum at a slower pace than their peers, and if *Data Analysis, Statistics, and Probability* content appeared later in the curriculum, then LEPs may not have been exposed to the material before they took the test in the spring. Future studies should investigate this hypothesis. Some content experts suspected that the use of the words *yard, feet, quart,* and *ounce* can be difficult for LEPs who are accustomed to the metric system, but none of the items that contained these words showed serious DIF. This finding again points to a possible effect of educational exposure.

The math test considered in this study was not very linguistically complex. First of all, a unidmensional IRT model fit the items as well as the two-dimensional model, suggesting that the effect of an extra language dimension was minimal. In addition, most of the items received ratings of one or two in lexical and grammatical complexity. The regressions of threshold differences or MH DIF effect sizes on linguistic predictors showed that LEPs did not find items more or less difficult as a function of linguistic complexity. However, higher lexical complexity and a greater number of words tended to make an item a less sensitive indicator of math ability for LEPs relative to non-ELs. In addition, higher grammatical complexity controlling for lexical complexity and the total number of words in an item tended to make the item less discriminating for non-ELs. The finding that language variables were not found to be a significant source of DIF with respect to thresholds implies that the particular test considered in this study was constructed carefully. However, this study also showed that language variables such as lexical complexity and wordiness can be a potential cause of smaller discrimination parameters for LEPs.

Suggested future research is obtaining English proficiency data for LEP students, and dividing the LEP population into groups with high, middle and low English proficiency and

comparing each one to non-ELs could provide further insight into the language variables that test constructors should take into account. One limitation of this study was the small number of linguistic experts. Obtaining ratings from a larger number of experts would improve the reliability of the findings in the current study. In addition, think-aloud protocols asking students to talk about their thoughts while they solve the math problems would give insights to the sources of DIF.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. Educational Assessment, 8, 231-257.

Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2005). The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives (CSE Rep. No. 663.) Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. Educational Measurement: Issues and Practice, 25, 36-46.

Abedi, J., & Herman, J. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. Teachers College Record, 112, 723-746.

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). NAEP math performance and test accommodations: Interactions with student language background (CSE Tech. Rep. No. 536). Los Angeles: University of California, National Center for Research on Evaluations, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. Applied Measurement in Education, 14, 219-234.

Abedi, J., Lord, C., & Hofstetter, C. (1998). Impact of selected background variables on students' NAEP math performance (CSE Tech Rep. NO. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Plummer, J. (1997). Language background as a variable in NAEP mathematics performance (CSE Tech Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). Standards for Educational and Psychological Testing.

Beal, C., Adams, N. M., & Cohen, P.R. (2010). Reading proficiency and mathematics problem solving by high school English language learners. Urban Education, 45, 58-74.

Butler, F., Bailey, A., Stevens, R., Huang, B., & Lord, C. (2004). Academic English in fifth-grade mathematics, science, and social studies textbooks (CSE Tech Rep. No. 642). Los Angeles, CA: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Callahan, R., Wilkinson, L., & Muller, C. (2010). Academic achievement and course taking among language minority in U.S. schools: Effects of ESL placement. Educational Evaluation and Policy Analysis, 32, 84-117.

Cummins, D.D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. Cognitive Psychology, 20, 405-438.

Eid, G. (2002). Gender, ethnicity, and language influences on differential item functioning in the SAT. Dissertation presented to Ohio University.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Johnson, E., & Monroe, B. (2004). Simplified language as an accommodation on math tests. Assessment for Effective Intervention, 29, 35-45.

Larsen, S.C., Parker, R.M., & Trenholme, B. (1978). The effects of syntactic complexity upon arithmetic performance. Learning Disabilities Quarterly, 1, 80-85.

Lee, M.K., Wells, C.S., & Sireci, S.G. (2011). Assessing Measurement Invariance in the Context of Disparate Sample Sizes and Proficiency Distributions. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the National Assessment of Educational Progress. International Journal of Testing, 8, 14-33.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. Harvard Educational Review, 78, 333-368.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners with math tests. Educational Assessment, 14, 160-179.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Ockey, G. J. (2007). Investigating the validity of math word problems for English language learners with DIF. Language Assessment Quarterly, 4, 149-164.

Raudenbush, S.W., Fotiu, R.P., & Cheong, Y.F. (1998). Inequality of access to educational resources: A national report card for eighth-grade math. *Educational Evaluation and Policy Analysis, 20,* 253-267.

SchleppegrEL, M.J. (2001). Linguistic features of the language of schooling. *Linguistics and Education, 12,* 431-459.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. Educational Assessment, 11. 105-126.

Solano-Flores, G., & Trunbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32,* 3-13.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. Psychological Methods, 11(4), 402-415.

Walker, C.M., Zhang, B., & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. Applied Measurement in Education, 21, 162-181.

Wheeler, L.J., & McNutt, G. (1983). The effects of syntax on low-achieving students' abilities to solve mathematics word problems. Journal of Special Education, 17, 309-315.

Wolf, M.K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. Educational Assessment, 14. 139-159.

Table 1. Descriptive statistics of the distribution of sum scores for non-EL, LEP, and former LEP students in the math test

	N	Mean	S.D.	Min	Median	Max	skewness
Non-ELs	Over 60,000	22.56	7.86	0	24	34	-0.42
LEPs	2,844	14	7.29	1	12	34	0.80
Former LEPs	1,314	17.9	8.03	2	17	34	0.24

Table 2. The content representation of the statewide grade 8 mathematics items used in this study

Content Strand	Percentage	Total number of items
Number Sense and Operations	29.4	10
Patterns, Relations, and Algebra	32.4	11
Geometry	8.8	3
Measurement	8.8	3
Data Analysis, Statistics and Probability	20.6	7

Table 3. Results of the DIF Analysis

Item	Content Strand	$\tau_{NonEL} - \tau_{LEP}$	$\lambda_{NonEL} - \lambda_{LEP}$	$\tau_{NonEL} - \tau_{FLEP}$	$\lambda_{NonEL} - \lambda_{FLEP}$	MH Category
11	Number Sense and Operations	-.27		-.22		B
12	Data Analysis, Statistics, and Probability	.40		.19		C (B)
17	Number Sense and Operations		-.09		-.10	
20	Patterns, Relations, and Algebra	.23				B
21	Patterns, Relations, and Algebra		-.09			
22	Number Sense and Operations		.11			
24	Data Analysis, Statistics, and Probability	.22				
29	Patterns, Relations, and Algebra		.11		.10	
31	Data Analysis, Statistics, and Probability	.26	.11			B
32	Data Analysis, Statistics, and Probability		.15			
34	Data Analysis, Statistics, and Probability	.19	.11			

Note. τ stands for the threshold parameter and λ for the factor loading. The alphabets in the last column represent the DIF categories for the non-EL vs. LEP comparison, and the one in the parenthesis represent the DIF category for the non-EL vs. FLEP comparison.

Table 4. Descriptive statistics of linguistic complexity of items

	Min	Max	Median	Mean	SD
Number of mathematical words	0	7	2	2.82	1.55
Lexical complexity rating	1	4	1	1.54	0.7
Number of general academic words	0	3	0.5	0.67	0.81
Number of low-frequency words	0	3	0	0.44	0.79
Grammatical Complexity rating	1	4	1	1.55	0.72
Total number of words	8	81	28.5	29.94	18.24
Total number of sentences	1	5	2	2.21	1.20
Number of passive voice phrases	0	2	0	0.12	0.41
Number of conditional / relative clauses	0	1	0	0.35	0.49
Number of modals	0	2	0	0.17	0.46
Number of nominalizations	0	1	0	0.06	0.24

Table 5. Regression models predicting the differences in thresholds and factor loadings between the reference group and the focal group.

	Model 1	Model 2	Model 3	Model 4
	MH DIF effect size in comparing non-ELs and LEPs	Differences in thresholds between non-ELs and LEPs	Differences in factor loadings between non-ELs and LEPs (34 items)	Differences in factor loadings between non-ELs and LEPs (items 18, 28, 31 removed)
Intercept	-0.01	-0.04	-0.02	-0.04
Lexical complexity	0.26	-0.05	0.05	0.06 *
Grammatical complexity	-0.11	0.03	-0.11 * (-0.20, -0.02)	-0.12 **
Total number of words	-0.01	0	0.004 ** (0.001, 0.006)	0.005 ***
Total number of sentences	0	0.05		
Number of general academic vocabulary words	0.14	-0.02		
Number of low-frequency vocabulary words	0.07	-0.01		
R^2	0.10	0.11	0.38	0.67
$\hat{\sigma}$	0.77 (27 *df*)	0.15 (27 *df*)	0.06 (30 *df*)	0.04 (27 *df*)

'*' denotes $p < .01$, '**' denotes $p < .001$, '***' denotes $p < .0001$.

Figure 1. Histograms of the math scores from three groups

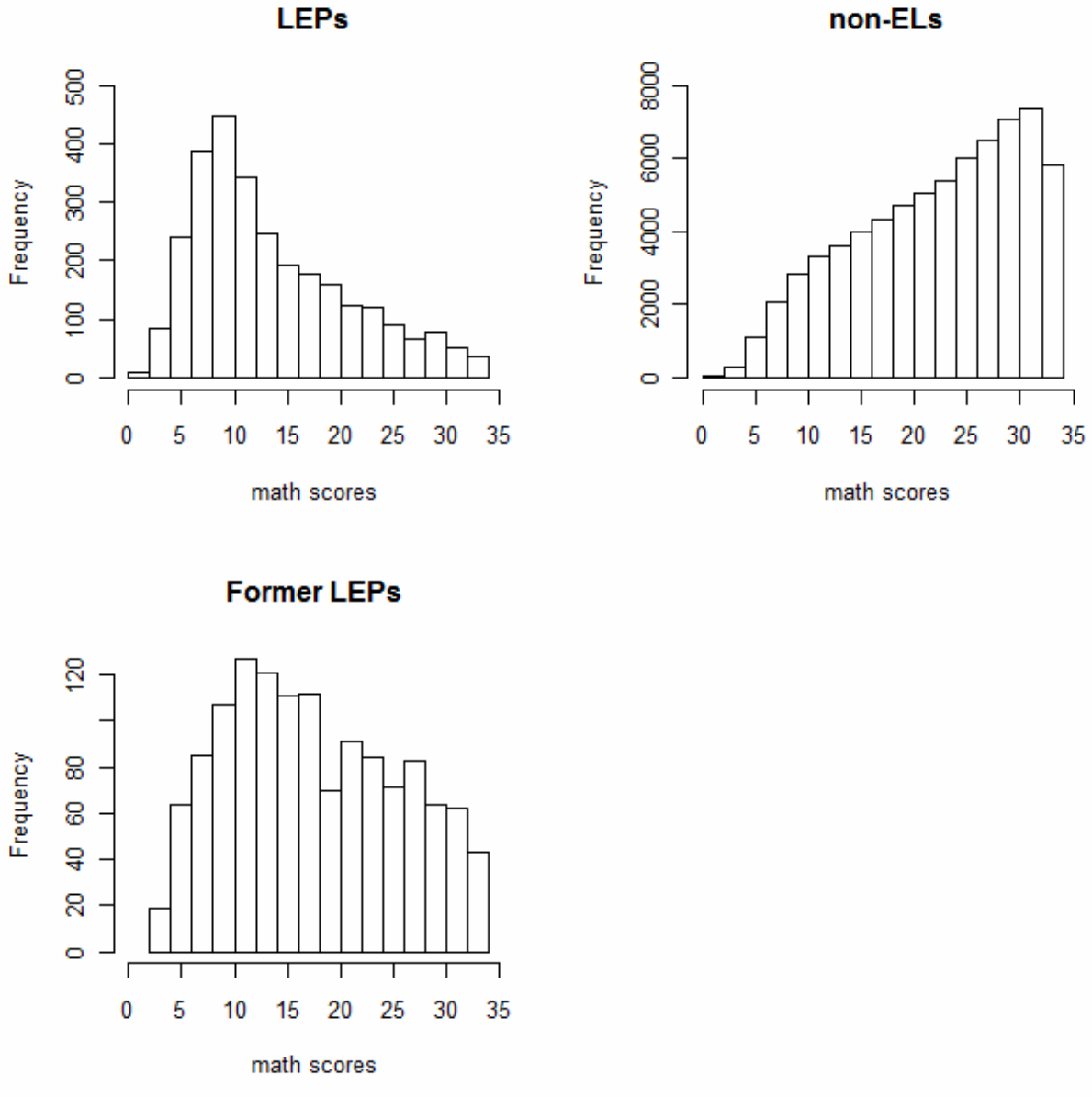


Figure 2. Histograms of the differences in thresholds and factor loadings

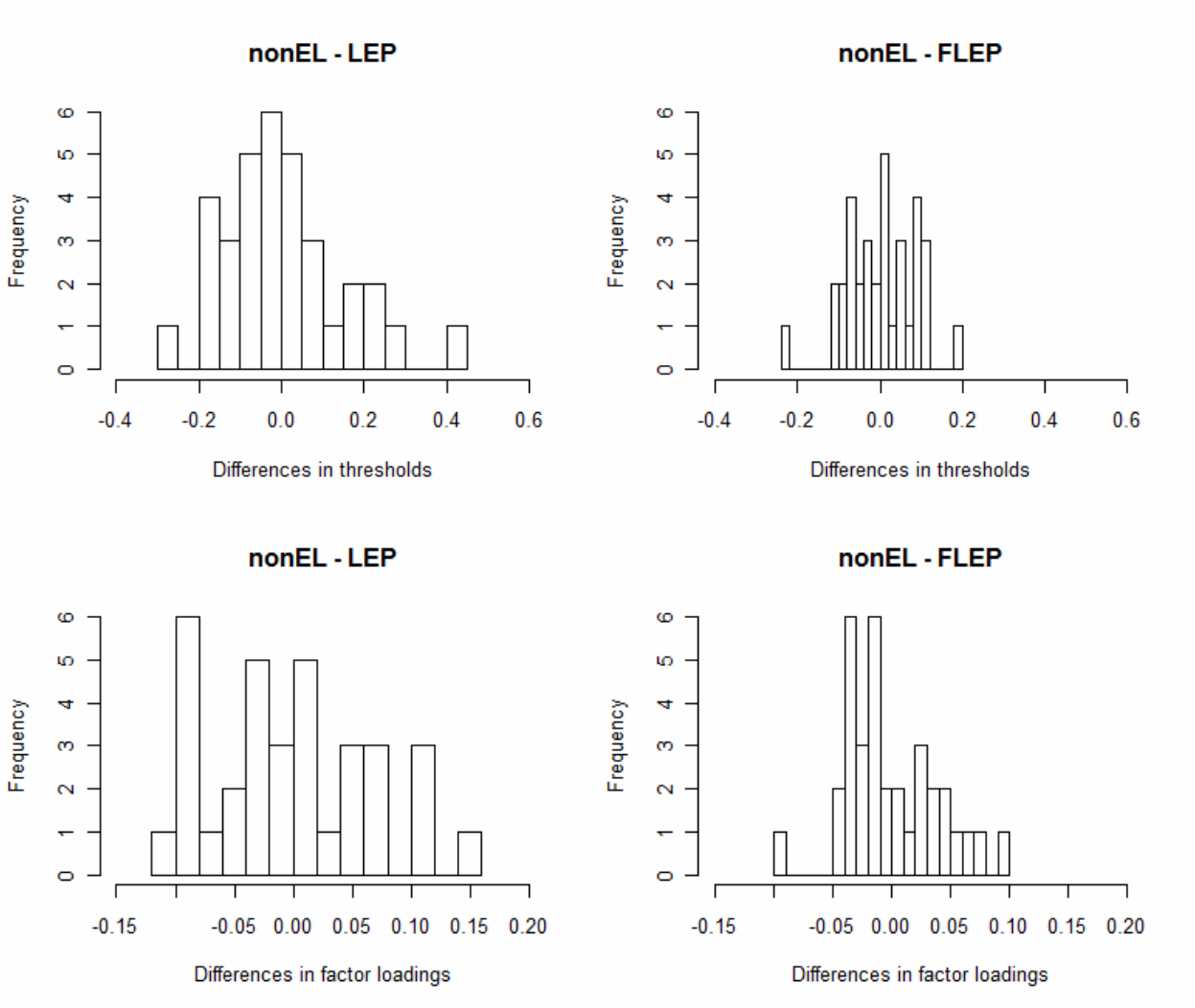


Figure 3. Thresholds in the LEP group plotted against the thresholds in the non-EL group. A line through the origin with a slope of 1 has been superimposed.

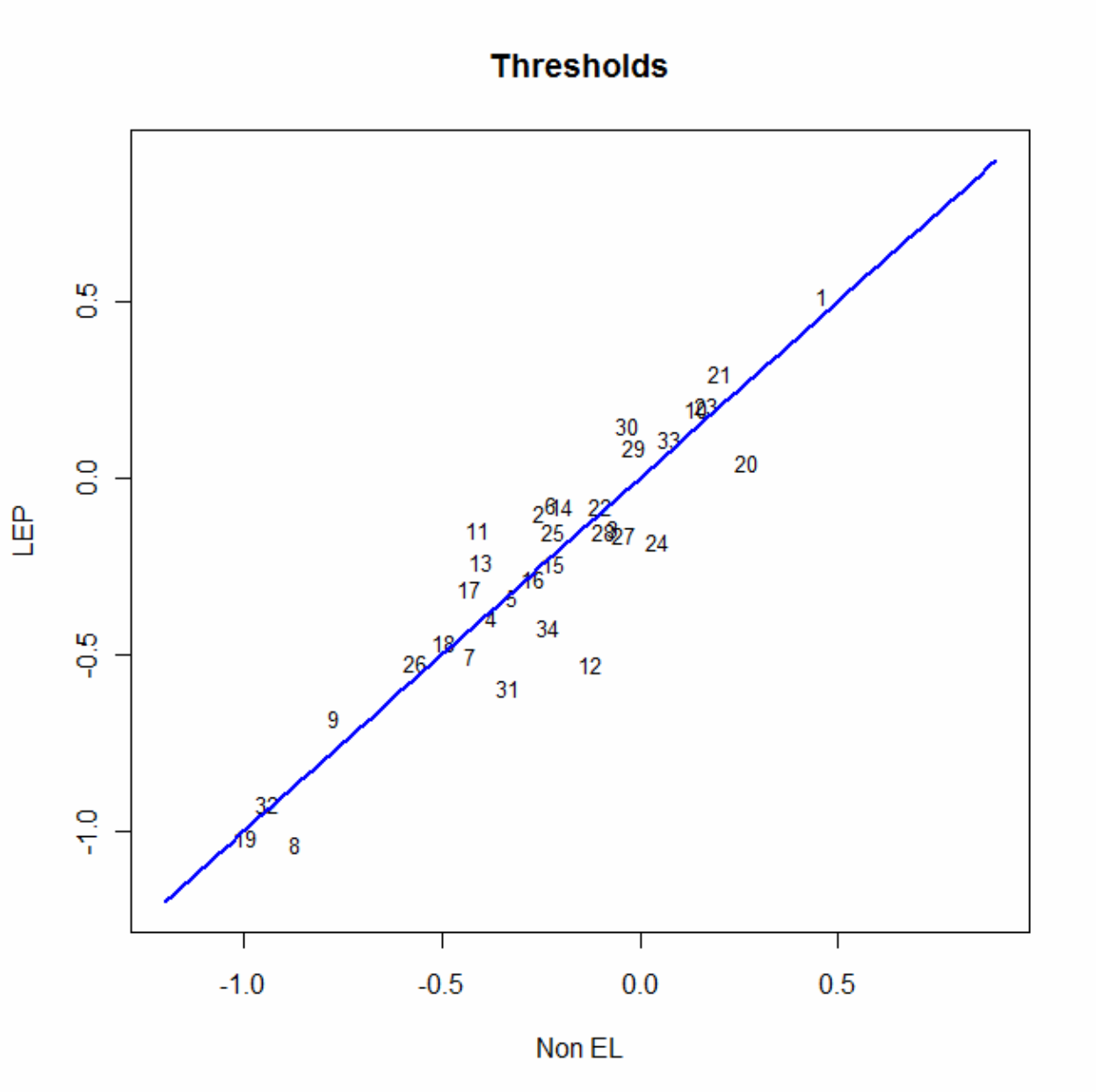


Figure 4. Thresholds in the FLEP group plotted against the thresholds in the non-EL group. A line through the origin with a slope of 1 has been superimposed.

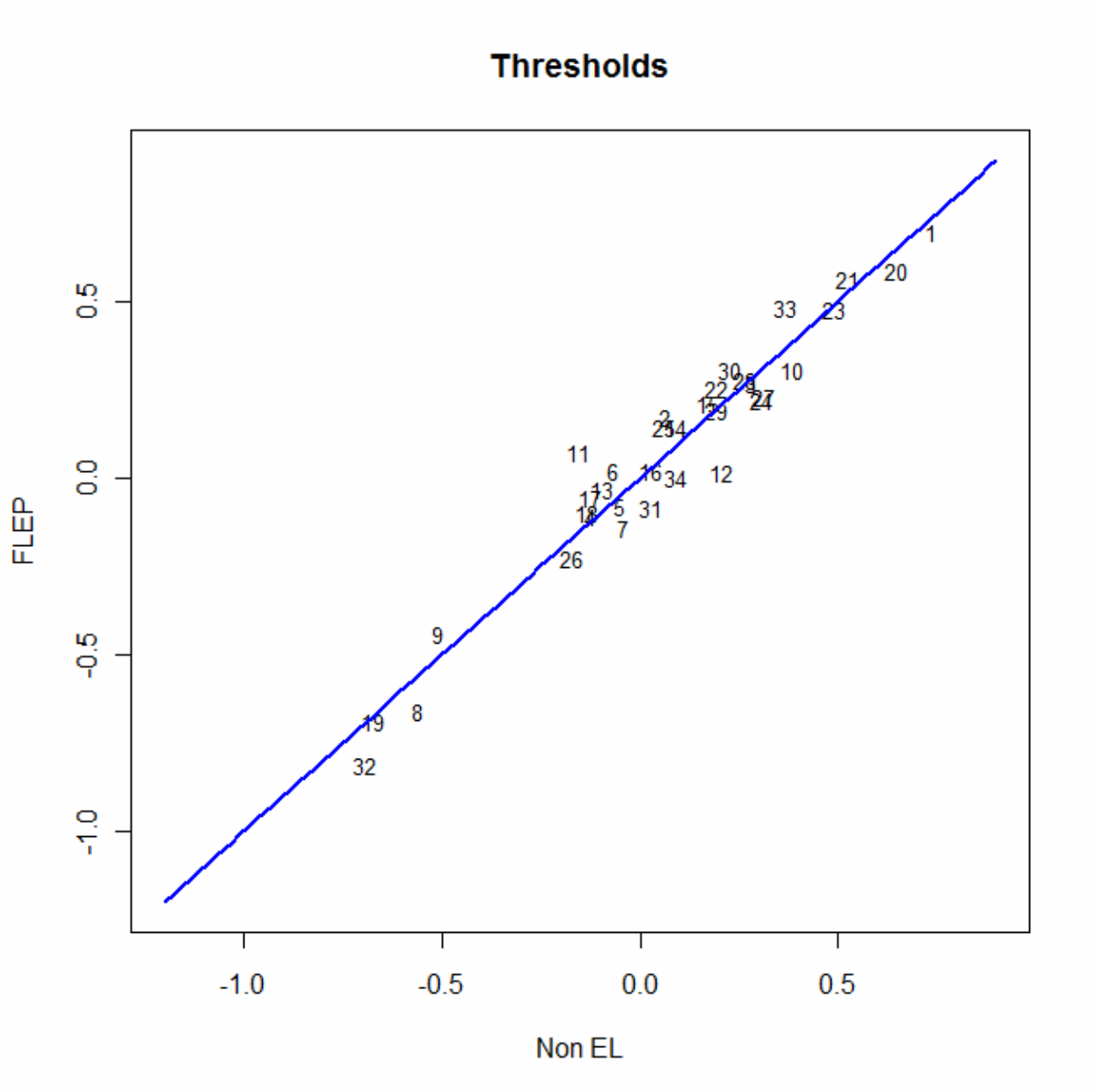


Figure 5. Factor loadings in the LEP group plotted against the factor loadings in the non-EL group. A line through the origin with a slope of 1 has been superimposed.

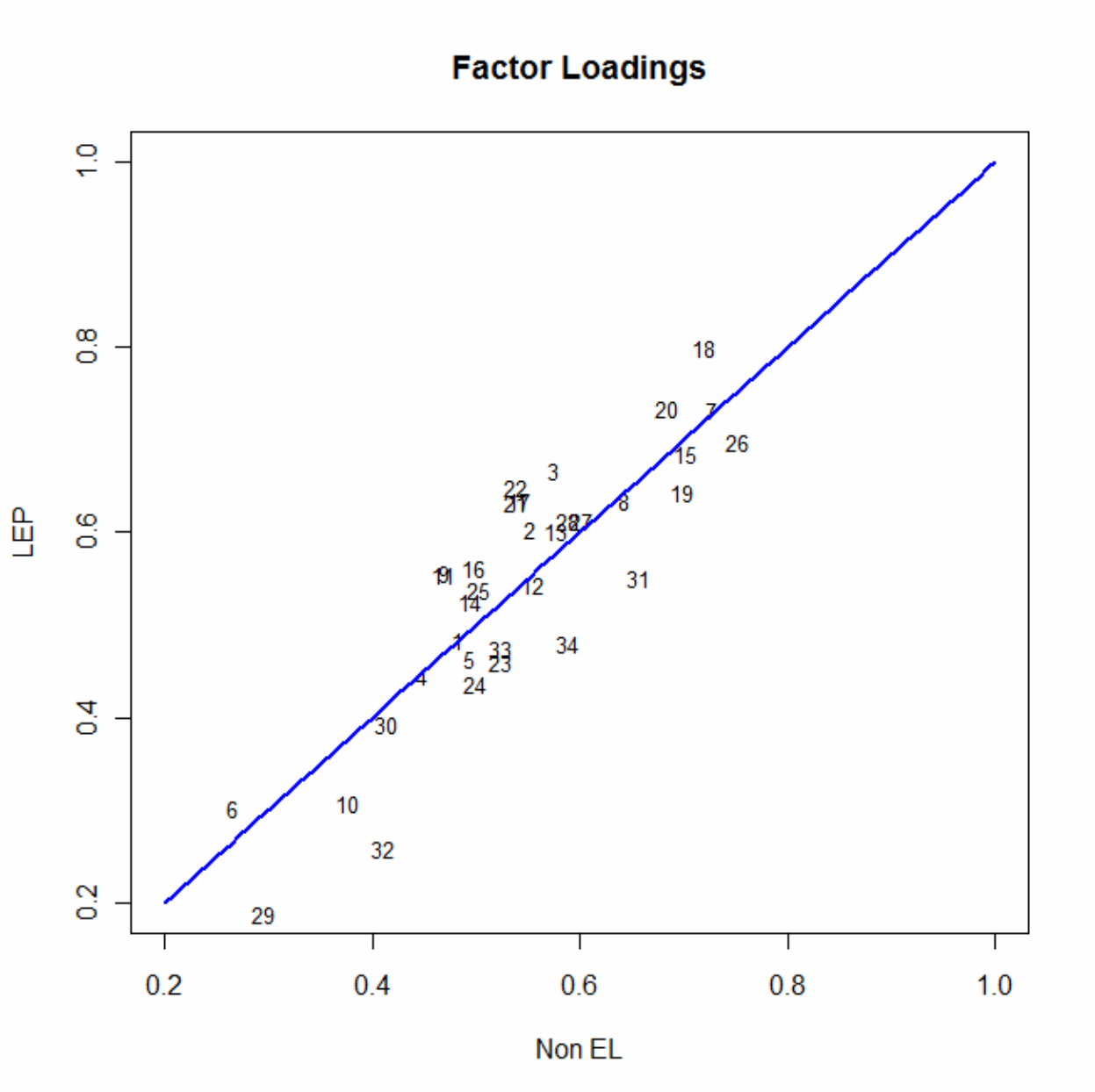


Figure 6. Factor loadings in the FLEP group plotted against the factor loadings in the non-EL group. A line through the origin with a slope of 1 has been superimposed.

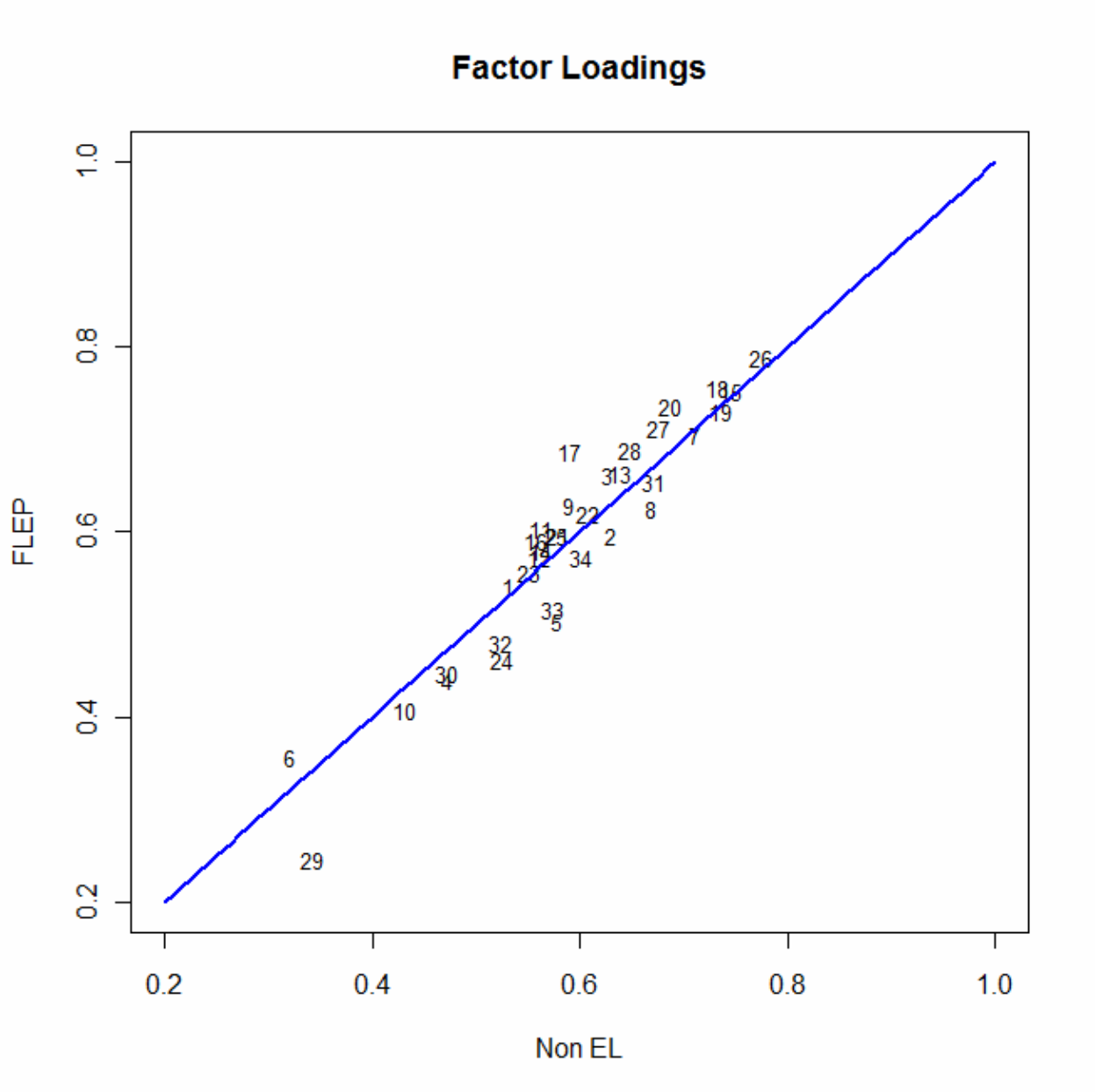


Figure 6. Test characteristic curves in each content area for non-ELs and LEPs

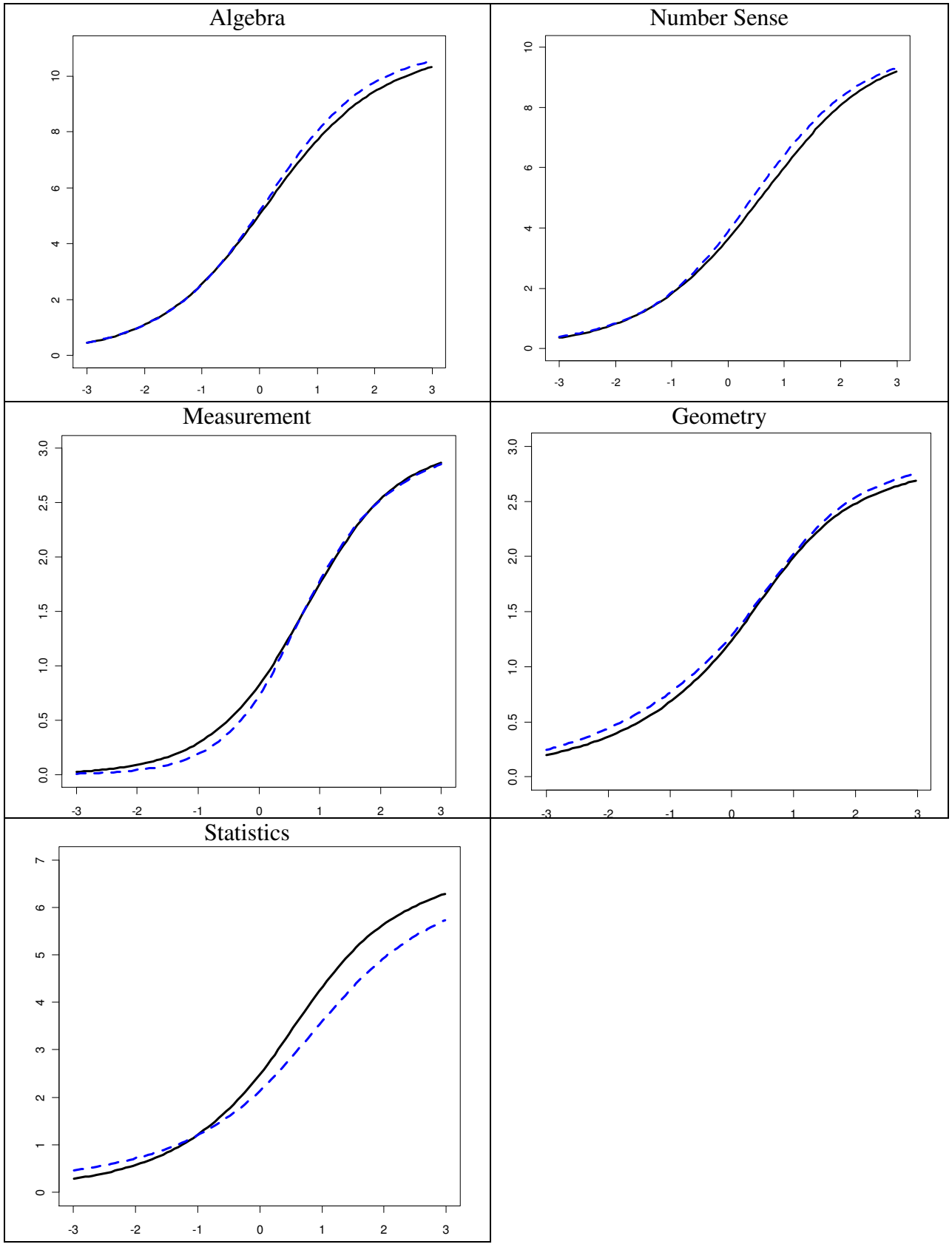


Figure 7. Test characteristic curves in each content area for non-ELs and FLEPs

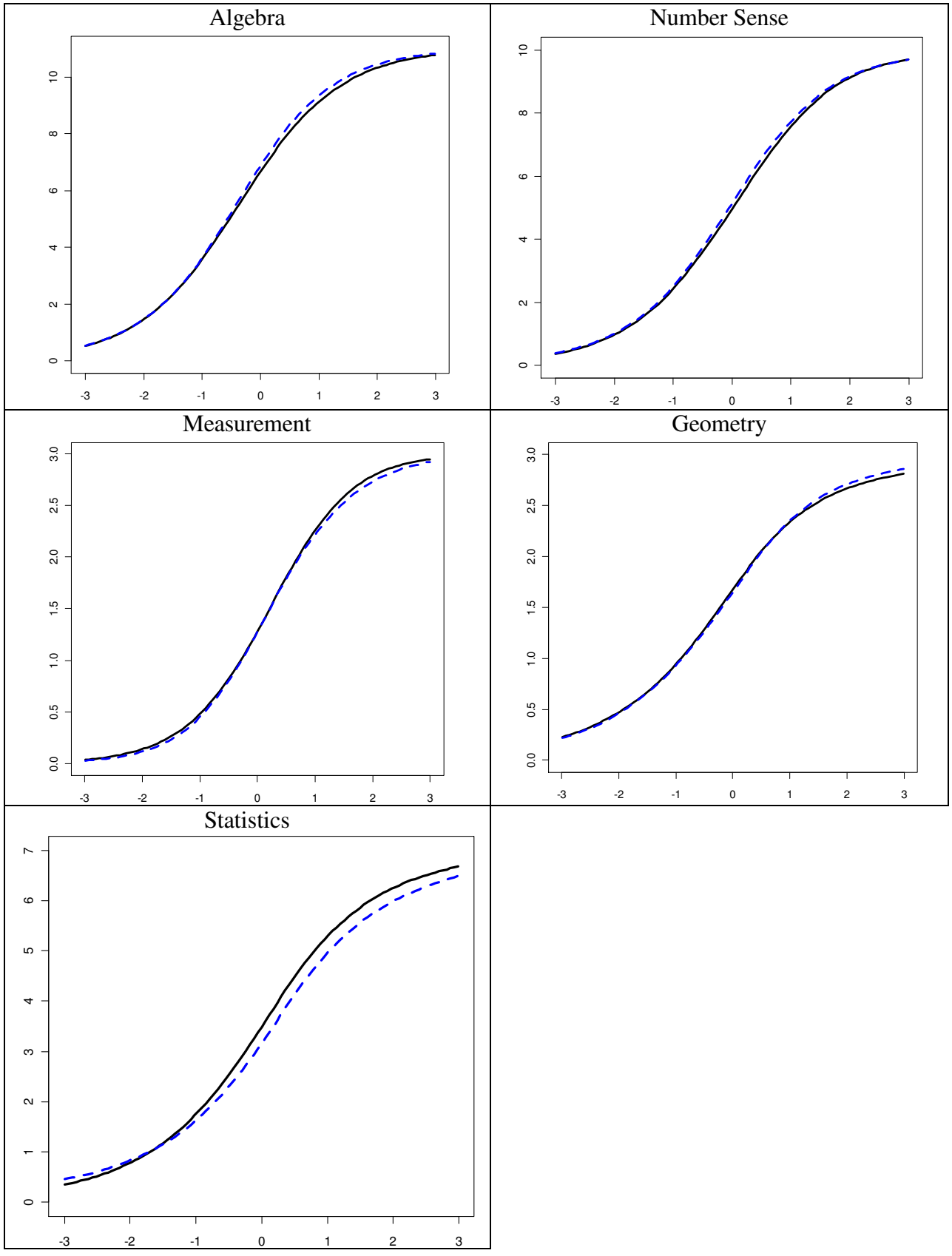


Figure 8. Histograms of grammatical complexity ratings for four raters

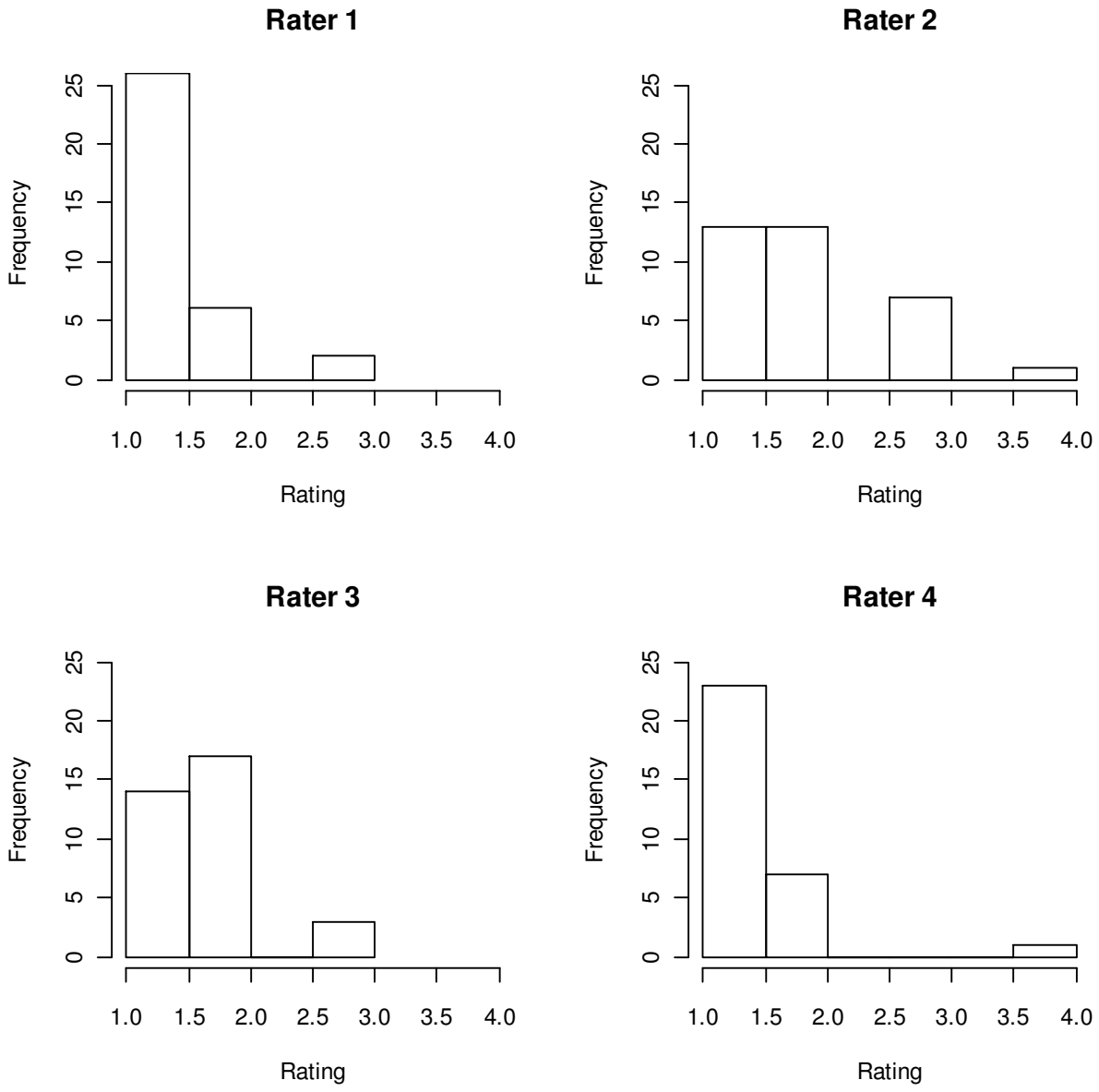


Figure 9. Histograms of lexical complexity ratings for four raters

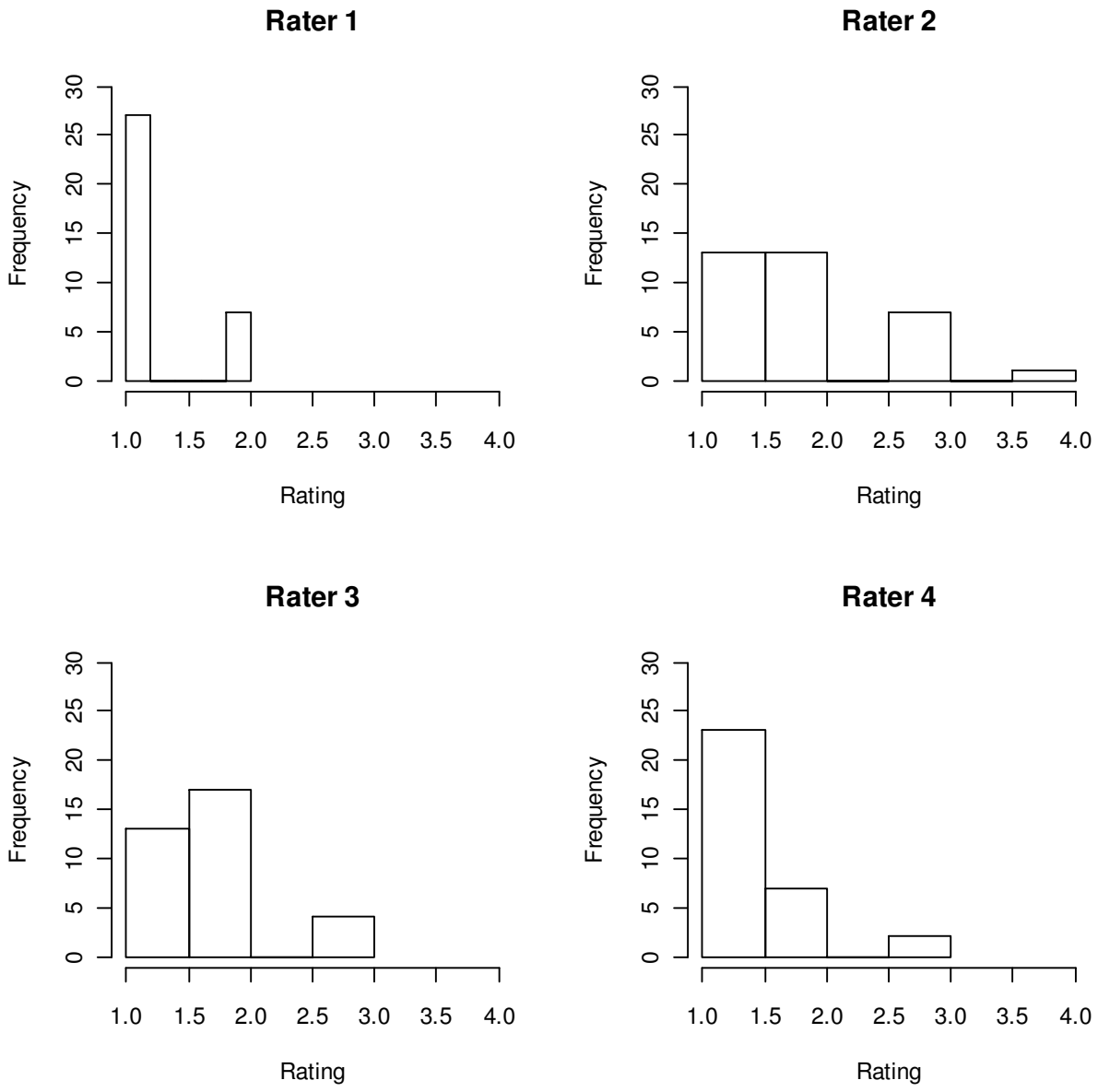
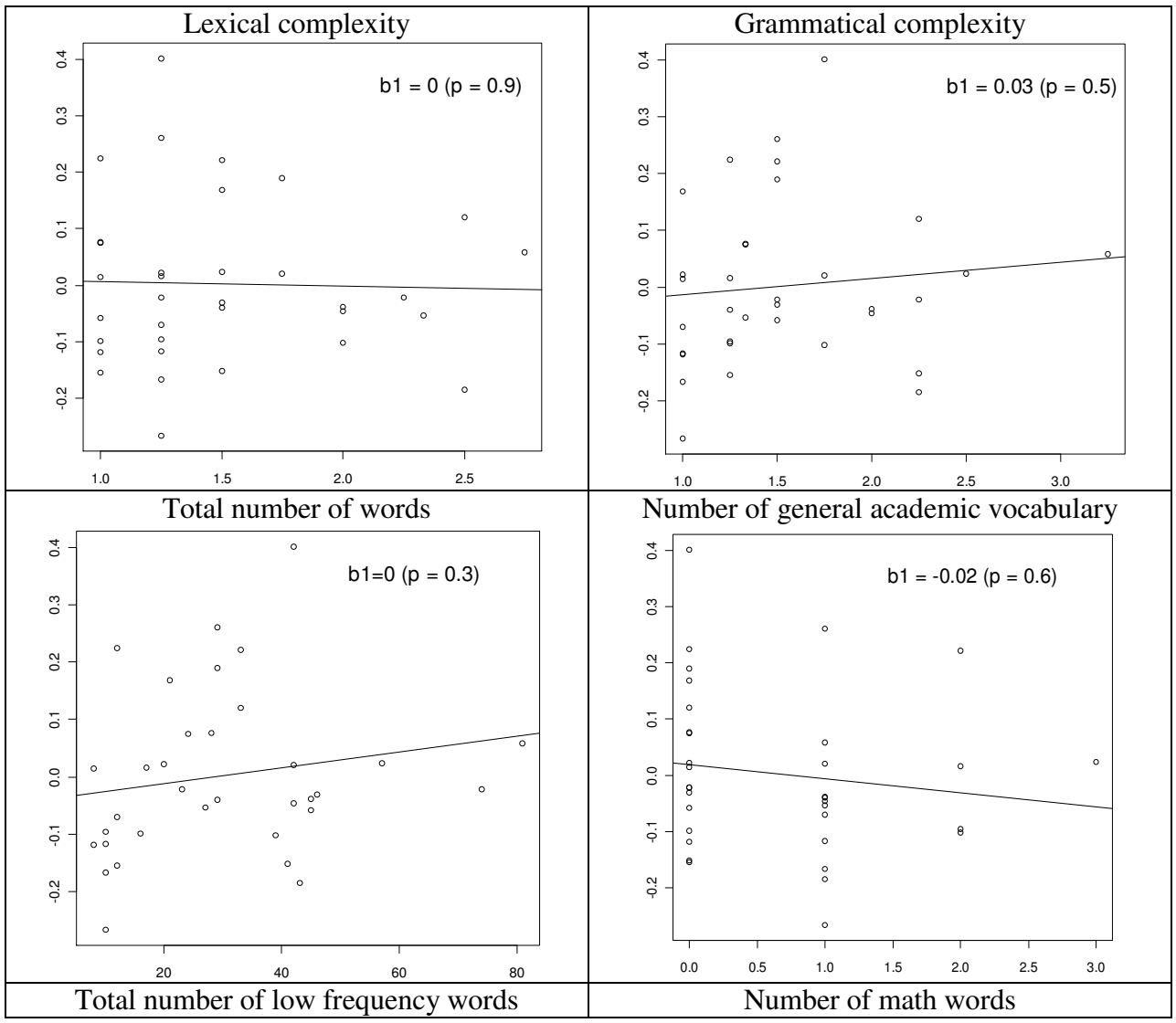


Figure 10. Regression of threshold DIF on predictors (non-ELs and LEP comparison)



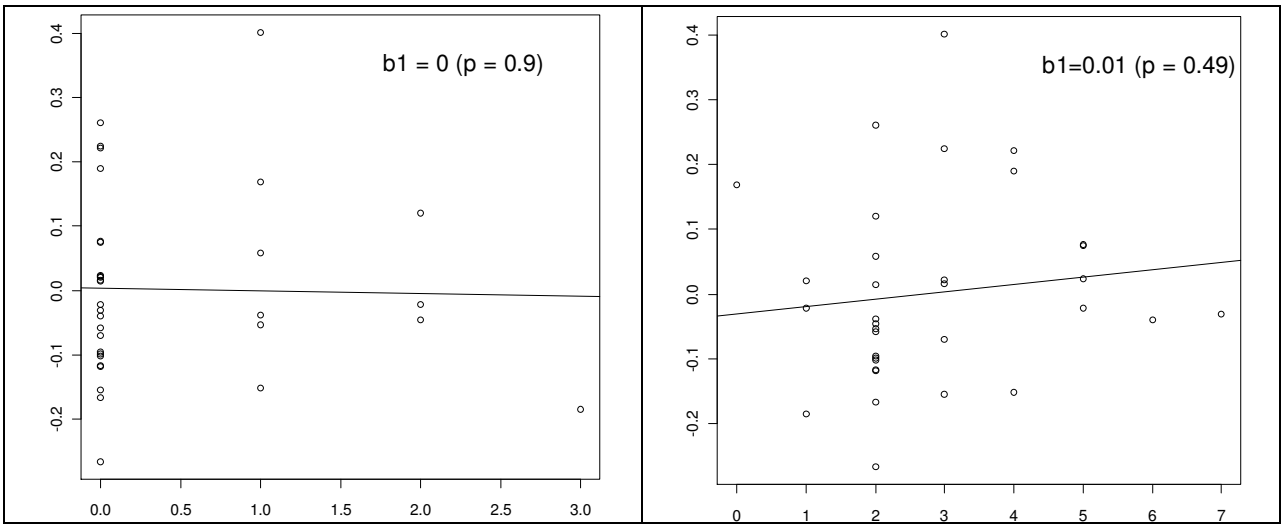
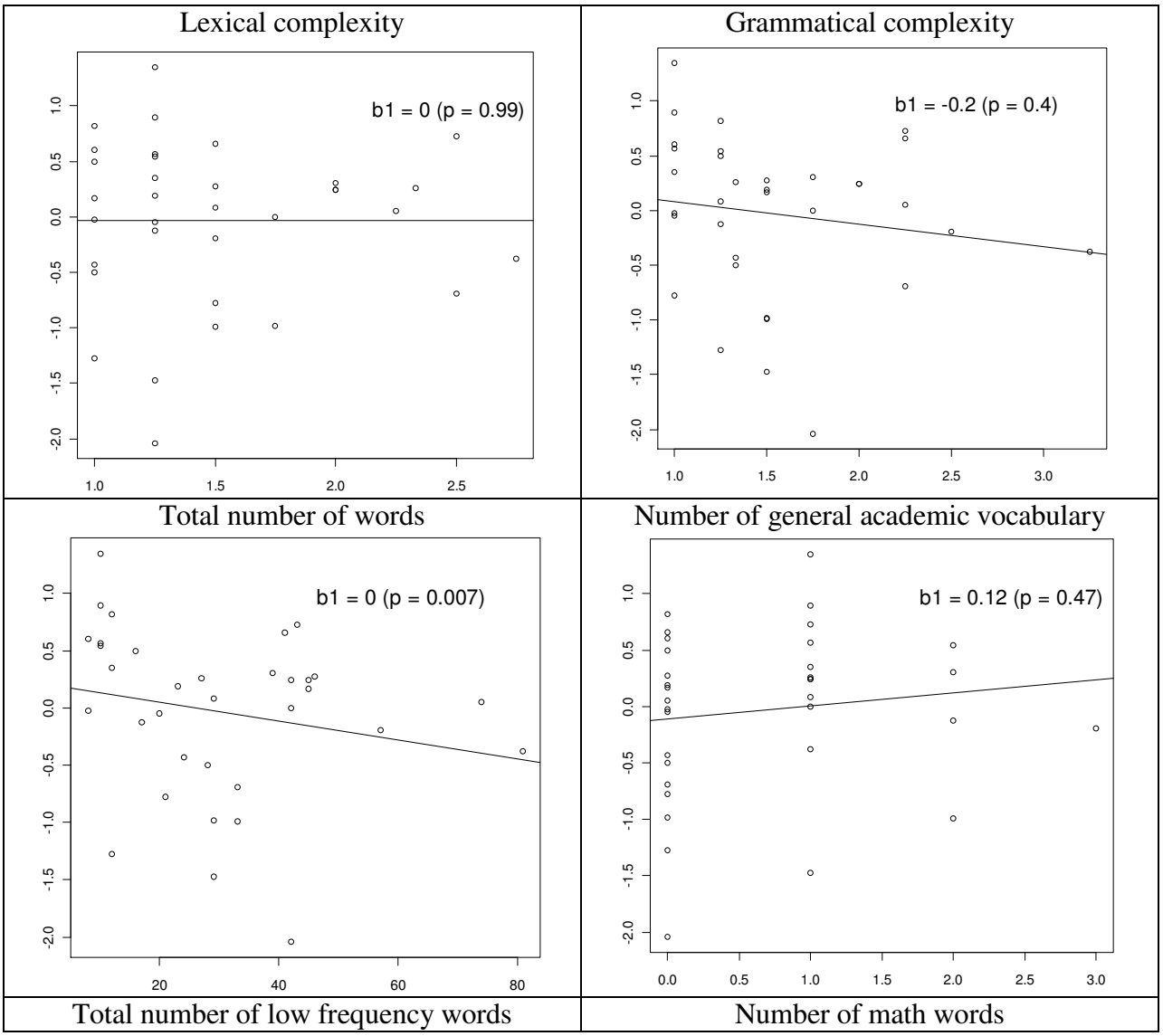


Figure 11. Regression of MH DIF effect size on predictors (non-ELs and LEP comparison)



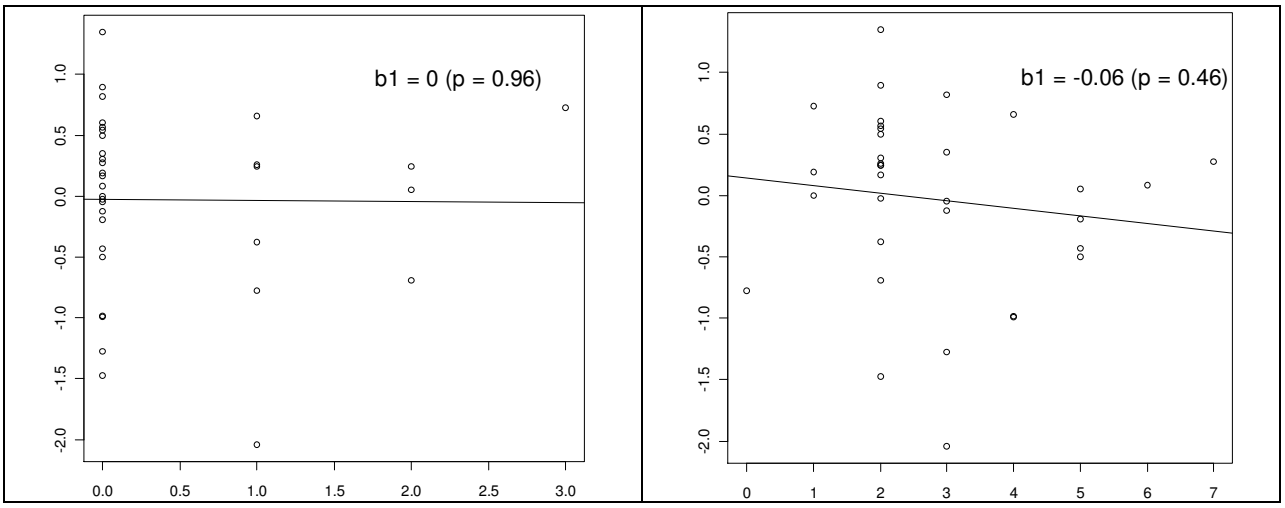
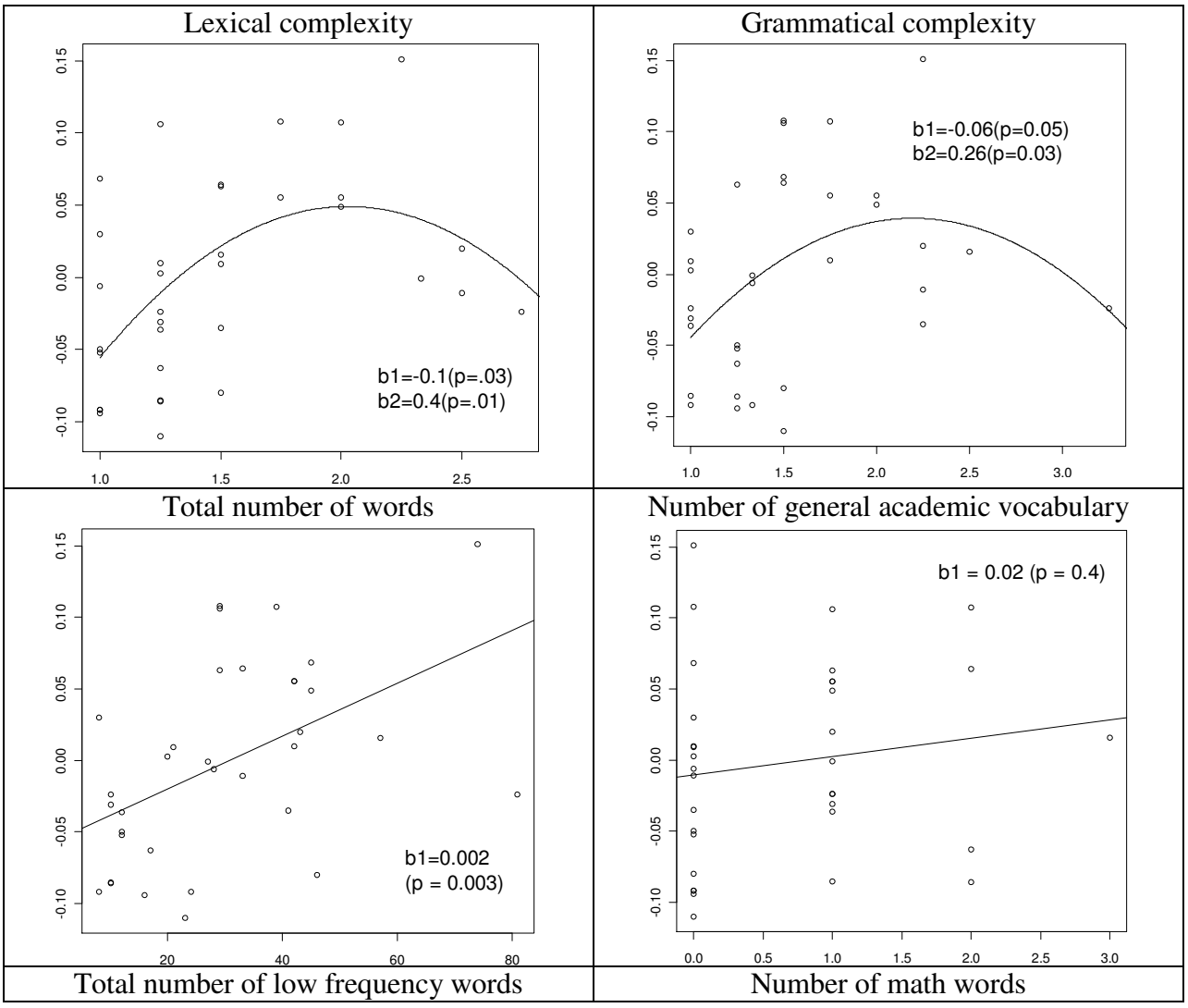
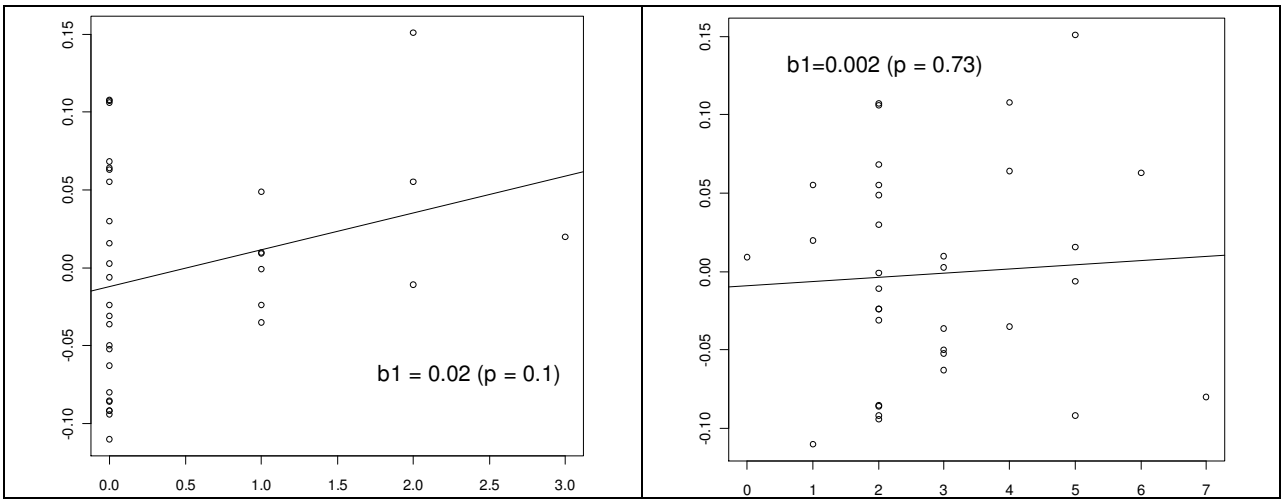


Figure 12. Regression of factor loading DIF on predictors (non-ELs and LEP comparison)




b1 = 0.02 (p = 0.1)
b1=0.002 (p = 0.73)