Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 28 Number 5, March 2023 ISSN 1531-7714

Evaluating the Equality of Regression Coefficients for Multiple Group Comparisons: A Case of English Learner Subgroups by Home Languages[1]

Hanwook Yoo, *Educational Testing Service*, Mikyung Kim Wolf, *Educational Testing Service*, and Laura D. Ballard, *Educational Testing Service*

As the theme of the 2022 annual meeting of the American Education Research Association, cultivating equitable education systems has gained renewed attention amid an increasingly diverse society. However, systemic inequalities persist for traditionally underserved student populations. As a way to better address diverse students' needs, it is of critical importance to understand different subgroups' performances. In the educational measurement field, evaluating the differences among multiple groups is an important consideration in addressing fairness issues for diverse groups of students. This article offers one technique to do so. It demonstrates how commonly-used multiple regression analysis can be applied to evaluate the equivalence of predictive structure across multiple groups in place of the factor analytic approach that requires a relatively large sample size per subgroup and strong assumptions. The technique is utilized in examining the relationship between English language proficiency and academic performance of English learners in one state when the subgroups are categorized by home language. The results showed statistically significant group differences between the reference group (Spanish-speaking ELs) and other focal groups (different home-language ELs) in various levels of comparisons (model fit, model structure, and individual predictor weights). The strengths and limitations of a proposed multiple group regression (MGR) approach are discussed in the educational research context.

Keywords: English Learner (EL), English Language Proficiency Assessment, Home Languages, Multiple Group Regression (MGR)

Introduction

Evaluating test-taker subgroup differences is a challenging issue not only methodologically (e.g., data availability, confounding variables, statistical techniques) but also conceptually (e.g., subgroup characteristics, interpretations). Yet, it has become increasingly important to cultivate diversity, equity, and inclusion in education systems (Barnett, 2020; Gay, 2018; Hays-Thomas, 2016). The underlying intent of current test-based accountability systems is to ensure that all students have the appropriate educational opportunities and improve their learning. However, test-based inferences and decisions can lead to differential impact on different groups (Lane, 2020). Lane (2020) stresses that examining these different impacts on subgroups (e.g., English learners, students

[1] The work presented here was supported in part by the William T. Grant Foundation grant ID187863 (An Investigation of the Language Demands in Standards, Assessments, and Curricular Materials for English Learners) to Educational Testing Service. The opinions expressed are those of the authors and do not represent views of the Foundation.

with disabilities, racial/ethnic subgroups, etc.) should be integral to making validity arguments and addressing fairness. To ensure fairness in the use of assessments for all students, Jonson and Geisinger (2022) also point out that the performances of traditionally underserved or marginalized groups of students need to be carefully examined first. For example, in the case of the English learner (EL) subgroup, English language proficiency (ELP) assessments play a substantial role in states' educational and accountability systems in addition to academic achievement assessments in content areas (e.g., language arts/reading, mathematics, science). ELP assessments are used not only to measure EL students' current proficiency level and progress of ELP attainment for accountability purposes, but they are also used to help determine whether EL students are adequately proficient to exit EL services. Despite the tremendously heterogeneous characteristics of the EL subgroup, validity investigations of ELP assessments have typically been conducted with the EL subgroup as a whole. One of the challenges in investigating differential impact on traditionally underserved students, such as EL students, is to define the subgroups and then deal with what is often a small sample size in each group, which makes robust interpretations of the results difficult.

Previous researchers have used statistical approaches to evaluate subgroup differences using null hypothesis significance testing (NHST; Keselman et al., 1998). Analysis of variance (ANOVA) is one of the commonly-applied statistical methods to demonstrate the equivalence of outcomes across different grouping variables (Cardinal & Aitken, 2006). However, testing equivalence of multiple group comparisons based on NHST (e.g., evaluating mean differences between groups through t-tests or ANOVA) is criticized because statistically significant results will not provide sufficient evidence to support that compared groups are equal (or comparable) to each other (Rusticus & Lovato, 2011; Tryon, 2001).

Another general method of evaluating group differences is using multi-group models in structural equation modeling (SEM; Jöreskog, 1971; Sorböm, 1974). Multiple group confirmatory factor analysis (MG-CFA; testing the equivalence of latent factor means) and MG-SEM (testing the equivalence of a causal structure) are the most popular approaches in the context of measurement invariance research (Byrne, 2012). These methods generally apply the same structural model to each group and evaluate the estimates of within-group parameters such as factor loadings, paths, and correlations. The goodness of model fit indices from the overall multi-group model are compared to gradually constrained models (i.e., weak, strong, strict, and structural invariance). For example, two nested models, a model with a set of parameters constrained to be equal across subgroups and a model with those same parameters freely estimated within each subgroup, are compared using likelihood ratio testing.

Although SEM-based methods are valuable for evaluating group differences, the processes and details can also become exceedingly complex because SEM is basically designed to analyze the relationships between latent variables. In addition, SEM requires several major assumptions to ensure accurate inferences, such as multivariate normality, no systematic missing data, correct model specification, and a sufficiently large sample size (Kline, 2012). There is no consensus about the appropriate sample size in SEM, but the rule of thumb is 100 cases or more per group for multi-group modeling (Kline, 2015). The SEM's sample size requirement, which is relatively large compared to other statistical techniques, can be more difficult to conduct with a group comparison analysis since increasingly specific subgroups with smaller sample sizes are requested these days (e.g., intersectional differential item functioning; Russell, Szendey, & Li, 2022).

Hence, in lieu of conducting factor-analytic multi-group comparison analyses that may require a large sample size per group, we propose a more practical method of evaluating the equality of regression coefficients across subgroups. Other social science studies have widely applied the idea of testing the equality of regression coefficients (Meng, Rosenthal, & Rubin, 1992; Paternoster, Brame, Mazerolle, & Piquero, 1998; Steiger, 1980). However, educational research has not widely applied this approach to compare group differences. In the present study, we aim to introduce a multiple group regression (MGR) analysis to examine the equivalence of predictive relationships in the regression model across multiple groups. We first describe the three steps of MGR analysis using different significance tests. Then, we demonstrate how this method can be used to analyze the relationship between English language proficiency

and academic performance of English learners in one state when the subgroups are categorized by home language. We also discuss the strengths and limitations of the MGR approach in the context of educational research.

Evaluating the Equality of Regression Coefficients across Multiple Groups

A key question of MGR approach is whether predictors from a multiple regression model explain the predictive relationship equivalently across subgroups within a heterogeneous population. The outcomes of regression analysis from each subgroup are compared to answer the following three sub-questions:

1. Do the hypothesized predictors provide a better prediction for one group than another ($H_0: \rho_r = \rho_f$)?
2. Are the models for different subgroups interchangeable ($H_0: \rho_{yr} = \rho_{yf}$)?
3. Are the regression weights of each predictor statistically equivalent across subgroups ($H_0: b_{ir} = b_{if}$)?

The three steps to compare the multiple regression models across multiple groups[2] are as follows: First, we conducted the multiple regression analysis to estimate the hypothesized relationship between predictors and outcome measures for each subgroup separately, allowing the model parameter estimates to differ across subgroups. The prediction quality (R^2) of the regression model from the reference and focal group(s) are compared by Fisher's z-test (1921), where ρ_r and ρ_f are R^2 of reference and focal group, respectively; n_r and n_f are the sample size of reference and focal group, respectively.

$$Z = \frac{\rho_r - \rho_f}{\sqrt{\frac{1}{(n_r-3)} + \frac{1}{(n_f-3)}}} \quad (1)$$

Second, we compared the structures of the regression models from the reference and focal group(s) using Hotelling's t-test (1940) and Steiger's z-test (1980). Hotelling (1940) initially proposed this statistical test to compare two dependent correlation coefficients, such as non-nested regression models from the same population. This kind of statistical test was continuously modified and developed by other researchers (Dunn & Clark, 1969; Meng, Rosenthal, & Rubin, 1992; Steiger, 1980). Three different correlation values were compared; the difference of predicted criterion values weighted by reference group's regression model (e.g., Focal group's predicted score using reference group's regression model; crossed model; ρ_{yr}) and by corresponding focal group's regression model (e.g., Focal group's predicted score using focal group's regression model; direct model; ρ_{yf}), along with the correlation of predicted scores between two models (ρ_{rf}).

$$Hotelling's\ t = \frac{(\rho_{yr} - \rho_{yf})\sqrt{(n_f-3)(1+\rho_{rf})}}{\sqrt{2|R|}} \quad (2)$$

where

$$|R| = 1 + 2(\rho_{yr})(\rho_{yf})(\rho_{rf}) - \rho_{yr}^2 - \rho_{yf}^2 - \rho_{rf}^2$$

$$Steiger's\ z = \frac{(Z_{yr} - Z_{yf})\sqrt{n_f-3}}{\sqrt{2-2c}} \quad (3)$$

where

$$c = \frac{\rho_{rf}(1-2\bar{r}^2) - \frac{1}{2}\bar{r}^2\left(1-2\bar{r}^2-\rho_{rf}^2\right)}{(1-\bar{r}^2)^2}$$

and

$$\bar{r} = \frac{\rho_{yr}+\rho_{yf}}{2};$$

Note that Z_{yr} and Z_{yf} applied Fisher's Z transformation

Third, we evaluated each predictor's regression weight differences from two regression models by a z-test using pooled standard error values. The generic formula of the z-test is shown below, where b_{ir} and b_{if} are predictor i's regression weight of reference and focal group, respectively; $df_{b_{ir}}$ and $df_{b_{if}}$ are the degree of freedom of predictor i's regression weight of reference and focal group, respectively; and $SE_{b_{ir}}$ and

[2] Professor Calvin P. Garbin at University of Nebraska Lincoln provides good reference materials and statistical software to calculate Fisher's z, Hotelling's t, and Steiger's z (https://psych.unl.edu/psycrs/statpage/).

$SE_{b_{if}}$ are standard error of predictor *i*'s regression weight of reference and focal group, respectively.

$$z = \frac{b_{ir} - b_{if}}{\sqrt{\frac{\left(df_{b_{ir}} * SE_{b_{ir}}{}^2\right) + \left(df_{b_{if}} * SE_{b_{if}}{}^2\right)}{df_{b_{ir}} + df_{b_{if}}}}} \quad (4)$$

However, this standard error of predictor estimates is negatively biased for a relatively large sample (both groups n > 30). Clogg, Petkova, and Haritou (1995) suggested more rigorous approaches to this problem by computing standard errors and confidence intervals for the difference. We applied Brame, Paternost, Mazerolle, and Piquero's approach (1998) since the sample size of each group in this study is larger than 30.

$$z = \frac{b_{ir} - b_{if}}{\sqrt{\left(SE_{b_{ir}}{}^2 + SE_{b_{if}}{}^2\right)}} \quad (5)$$

Current Study: A Subgroup Analysis of the Relationship between English Language Proficiency and Academic Performance of English Learners

As a case of empirical demonstration, we examined the relationship between the statewide standardized content-area assessments in English language arts and mathematics (content assessment for simplicity, henceforth) and English language proficiency (ELP) assessment performances of English learner (EL) students from one state to better understand diverse groups of EL students' achievements and needs. Particularly, this study investigated whether the relationship varied across the subgroups of EL students by their home language (HL) background. Despite EL students' heterogeneous characteristics, no empirical research has examined EL students' performances on the state-wide assessments by their HL characteristics.

This line of research has provided important validity evidence for ELP assessment use, particularly in understanding the impact of ELP on EL students' demonstration of content knowledge and skills. That is, the findings would provide useful insight into the role of ELP assessments in understanding EL students' content performance, given the increased language demands of the new standards (Wolf et al., 2022). The findings also have implications for high-level EL instruction planning and EL reclassification decisions.

Data

We obtained one state's Grade 5 datasets from the 2018-2019 school year, including students' HL background, individualized education program (IEP), free/reduced lunch program participation (as an indicator of socio-economic status), gender, and scores from content and ELP assessments. In terms of the HL background, 43% and 28% of 7,439 EL students were reported as Spanish and Arabic as HLs, respectively. Other HL group sizes were smaller than 200 students in each HL group. In selecting different HL groups, we chose the language groups with a sample size of at least 80 students. Since there were considerably unbalanced sample sizes in Spanish-speaking (N=3,165) and Arabic-speaking (N=2,062) subgroups, we conducted stratified random sampling for these two groups, taking their content and ELP assessment performances and background information into consideration. Table 1 presents a summary of the final sample included in the analysis.

Among the seven largest HL groups, four HL groups (Spanish, Arabic, Bengali, and Albanian) were Indo-European languages, while the other three HLs (Vietnamese, Japanese, and Chinese) were Non-Indo-European languages. We hypothesized that there would be group differences because the reference group of this study was Spanish-speaking EL group (which is an Indo-European language). Note that the Japanese group had only 1% of EL students who participated in the free/reduced lunch program, while 91% of the Spanish group received free/reduced lunch.

EL students completed the State's content assessment, which included Smarter Balanced English language arts (ELA) and mathematics items. ELA test performance was reported as a total score and four subscores (i.e., reading, writing, listening, and research), while math test performance was reported as a total score and three subscores (i.e., concepts and procedures, problem solving and modeling & data analysis, and communicating reasoning). EL students also took the WIDA ELP assessment (i.e., ACCESS for ELLs), which evaluates the language demands needed to reach college and career readiness.

Table 1. EL Students' Background Information by Home Language

Home Language	N	Gender (%)		FRL (%)		IEP (%)	
		Male	Female	No	Yes	No	Yes
Total	7,439	54	46	18	82	87	13
Spanish (All)	3,165	53	47	11	89	85	15
Arabic (All)	2,062	55	45	8	92	90	10
Spanish (Sampled)	306	51	49	9	91	86	14
Arabic (Sampled)	210	60	40	8	92	90	10
Bengali	190	51	49	8	92	95	5
Albanian	132	61	39	24	76	82	18
Vietnamese	100	46	54	31	69	89	11
Japanese	82	59	41	99	1	96	4
Chinese	80	65	35	55	45	80	20

Note. FRL=Free/Reduced Lunch; IEP=Individual Education Program.

This ELP assessment measured English language ability in the needed to reach college and career language ability in the four language domains of Listening, Reading, Speaking, and Writing. Appendix A summarizes the content and ELP assessment performances by HL groups.

Analyses

The results of a preliminary multiple regression analysis using the total sample (N=7,439) indicated that the ELP Reading scores significantly predicted EL students' performances on both ELA and math assessments. We compared multiple regression analysis results from the seven largest HL groups based on the regression model below;

$$\text{ELA (or Math)} = b_0 + b_1 Gender + b_2 FRL + b_3 IEP + b_4 Listening + b_5 Reading + b_6 Speaking + b_7 Writing \quad (6)$$

First, we conducted multiple regression analyses for each HL group to predict the ELA or math assessment performance by ELP language domain scores after controlling three background variables (i.e., gender, IEP, FRL). We examined four assumptions of multiple regression analysis (linearity, reliability, homoscedasticity, and normality) for each subgroup suggested by Osborne and Waters (2002). The multiple regression results from Spanish group were chosen as a reference outcome to compare the results of other HL groups. Second, we used Fisher's z-test to compare the model fit of regression models (R^2) whether predictors from other HL groups were equally predicted compared to Spanish group. Third, we used Hotelling's t-test and Steiger's z-test to evaluate the structural differences between other HL groups' regression models against Spanish group's model. We compared the difference of predicted criterion values weighted by Spanish model (e.g., Arabic EL students' predicted ELA score using Spanish group's regression model) and corresponding HL group model (e.g., Arabic EL students' predicted ELA score using Arabic group's regression model), along with the correlation of predicted scores between two models. Lastly, we evaluated the interaction between individual predictor and HL group membership by comparing the individual predictor's regression weights from other HL groups against the weights from Spanish group. Each predictor's unstandardized estimates and standard error of estimate were used to conduct the significance test (Brame et al., 1998).

A strict evaluation criterion was proposed to overcome the criticism of 'NHST as an equivalence test for group comparisons' (Rusticus & Lovato, 2011). We agreed that a statistical non-significance of one of three null hypotheses (i.e., model fit, model structure, or individual predictor) might be insufficient to claim that the finding provides evidence for multiple group comparability. Thus, we defined the 'equivalent predictive relationship across groups' when the statistical results of all three steps were non-significant. This rigorous rule would avoid the limitation of the null hypothesis testing since the chance would be rare for the equivalent findings to be falsely identified by all three different null hypothesis testings in the MGR approach. Note that we do not examine whether two groups are exactly equal; the word 'equivalence' in this

study indicates that the group difference exists, but the magnitude is quite trivial and acceptable to ignore.

Results

The results of the multiple regression model for Spanish group (i.e., reference group) are summarized in Table 2. As expected, the results from the regression analyses (after testing multiple models with various combinations of controlling covariates) revealed that ELP Reading scores were always significant in predicting the outcome (i.e., the ELA and math assessment scores for EL students). Although ELP Listening scores were also significant, the unstandardized estimate was small in predicting the ELA and math assessment scores (.07 for ELA; .08 for math). Regarding the controlling covariates, Gender status (0=Male; 1=Female) was a significant controlling student background variable for math performance.

We conducted a series of regression analyses to a) examine the relationships between ELA and math assessment performances and four ELP language domain subscores, and b) compare the models derived from six focal groups. Table 3 provides the multiple regression weights for six other HL groups compared to Spanish groups by Fisher's z-test. Comparisons of the fit of the predicted ELA model from the Spanish and other HL groups revealed that there was no significant difference between the respective R^2 values except for Japanese (R^2=.715, Z=-2.050, p=.040) and Chinese (R^2=.761, Z=-2.823, p=.005) groups. For the Chinese group, the ELP Reading predictor (.45) only had significant regression weights. On the other hand, the model for Japanese group had ELP Listening (.16), Reading (.21), and Speaking (.13) as predictors having significant regression weights; still, ELP Reading had the largest contribution. Comparisons of the fit of the predicted math model from the Spanish and other HL groups revealed that the R^2 values from all HL groups did not show any significant differences.

We also conducted a comparison of the structure of the predicted ELA models across different HL groups by applying the model derived from the

Table 2. Multiple Regression Results of Spanish HL Group (N=306)

Variable	ELA (R^2=.56; Adjusted R^2=.55)				Mathematics (R^2=.43; Adjusted R^2=.42)			
	Estimate	*SE*	*T*	*p*	Estimate	*SE*	*T*	*p*
Intercept	1,283.20	12.11	105.97	.000	1,322.04	14.43	91.63	.000
Gender	1.10	1.56	.71	.480	-7.53	1.86	-4.05	.000**
FRL	1.55	2.33	.67	.507	-4.61	2.78	-1.66	.098
IEP	-.33	2.39	-.14	.889	-3.74	2.85	-1.31	.190
Listening	.07	.03	2.71	.007**	.08	.03	2.64	.009**
Reading	.42	.04	11.61	.000**	.34	.04	7.80	.000**
Speaking	.02	.02	.75	.456	.04	.03	1.405	.161
Writing	.01	.03	.36	.721	-.03	.03	-1.01	.314

Note. SE=standard error of estimate; FRL=Free/Reduced Lunch; IEP=Individual Education Program; *p<.05; **p<.01.

Table 3. Regression Model Fit Comparison by Fisher's z-test

HL Group	N	ELA			Mathematics		
		R^2	Fisher's z	*p*	R^2	Fisher's z	*p*
Arabic	210	.499	1.002	.316	.435	-.072	.943
Bengali	190	.504	.900	.368	.459	-.392	.695
Albanian	132	.555	.122	.903	.502	-.879	.379
Vietnamese	100	.585	-.271	.786	.499	-.758	.448
Japanese	82	.715	-2.050	.040*	.548	-1.234	.217
Chinese	80	.761	-2.823	.005**	.497	.031	.975

Note. Reference group was the Spanish HL group (N=306; ELA R^2=.564, F(7,298)=55.0, p<.001; Math R^2=.430, F(7,298)=32.1, p<.001); *p<.05; **p<.01.

Spanish group to the data from the other HL groups (See Table 4). Comparing the resulting "crossed" R^2 with the "direct" R^2 originally obtained from each HL group. The Japanese group showed that the direct R^2 and crossed R^2 were significantly different, indicating the apparent differential structure of the regression weights compared to the Spanish group on both predicted ELA and math models. Note that the Japanese subgroup differs from the Spanish subgroup in ways other than HL grouping (i.e., FRL and IEP). There was no significant model structure difference from other HL groups compared to Spanish group when the structures of predicted ELA and math models were compared. For predicted math models, the p-values of the other two Asian HL groups (Vietnamese and Chinese) were marginally non-significant but quite close to p =.05.

We conducted further analysis to evaluate the differences in individual predictor weights of regression models between Spanish group and other HL groups. Table 5 provides the individual predictor difference of the predicted ELA model between Spanish and Japanese groups. The results showed that ELP Reading and Speaking scores had significantly different regression weights (Z=2.797, p<.01 and Z=-2.161, p<.05, respectively). Although ELP Listening scores showed different regression weights between two groups (.07 from Spanish and .16 from Japanese), the difference accounted by standard errors was not significant (Z=-1.709, p=.087).

Table 6 provides the individual predictor difference of the predicted math model between Spanish and Vietnamese groups. The results showed that ELP Reading scores had significantly different regression weights (Z=2.158, p<.05). Even though ELP Listening scores showed different regression weights between two groups (.08 from Spanish and .19 from Vietnamese), the difference accounted by standard errors was not significant (Z=-1.835, p=.067). Except for these two comparisons, there were no significantly different regression weights between Spanish group and other HL groups in predicted ELA and math models (See Appendices B and C).

As hypothesized, we could not find any statistically-significant group differences in three steps of MGR analysis when we compared three HL groups of Indo-European language (Arabic, Bengali, and Albanian) to the reference group. All statistical results for testing the equality of regression coefficients were non-significant for these three HL groups. However, we found at least one or more statistically-significant regression coefficient differences among three steps of MGR analysis for three HL groups of Non-Indo-

Table 4. Regression Model Structure Comparison by Hotelling's t-test and Steiger's z-test

Predicted Score	Statistics	Arabic	Bengali	Albanian	Vietnamese	Japanese	Chinese
ELA	R_{yr}	.707	.710	.745	.765	.846	.872
	R_{yf}	.705	.694	.732	.744	.792	.856
	R_{rf}	.998	.978	.982	.973	.937	.981
	Hotelling's t	.643	1.481	1.167	1.382	2.533	1.471
	p	.521	.140	.246	.170	.013*	.145
	Steiger's z	.643	1.473	1.161	1.369	2.450	1.452
	p	.520	.141	.246	.171	.014*	.146
Math	R_{yr}	.660	.677	.708	.707	.740	.705
	R_{yf}	.648	.662	.683	.658	.545	.640
	R_{rf}	.982	.977	.964	.932	.736	.907
	Hotelling's t	1.211	1.299	1.498	1.849	3.515	1.863
	p	.227	.195	.137	.067	<.001**	.066
	Steiger's z	1.207	1.294	1.486	1.820	3.318	1.826
	p	.227	.196	.137	.069	<.001**	.068

Note. R_{yr}: the original R from the corresponding HL groups' multiple regression model (Direct R); R_{yf}: the weights from Spanish multiple regression model applied to each HL group (Crossed R); R_{rf}: the predicted score correlation between two models (i.e., Direct R & Crossed R); *p<.05; **p<.01.

Table 5. Comparison of Individual Predictor Weights (Predicted ELA model; Japanese HL)

Predictor	Spanish		Japanese		SE(B-diff)	*Z*	*p*
	Estimate	*SE*	Estimate	*SE*			
Listening	.069	.025	.164	.050	.056	-1.709	.087
Reading	.423	.036	.205	.069	.078	2.797	.005**
Speaking	.016	.021	.125	.046	.050	-2.161	.031*
Writing	.010	.028	.042	.072	.077	-.415	.678

Note. SE=standard error of estimate; SE(B-diff) = $\sqrt{(SE_{br}{}^2 + SE_{bf}{}^2)}$; *p<.05; **p<.01.

Table 6. Comparison of Individual Predictor Weights (Predicted Math model; Vietnamese HL)

Predictor	Spanish		Vietnamese		SE(B-diff)	*Z*	*p*
	Estimate	*SE*	Estimate	*SE*			
Listening	.080	.030	.188	.050	.059	-1.835	.067
Reading	.339	.043	.156	.073	.085	2.158	.031*
Speaking	.036	.025	.041	.041	.048	-.114	.909
Writing	-.033	.033	.055	.071	.079	-1.122	.262

Note. SE=standard error of estimate; SE(B-diff) = $\sqrt{(SE_{br}{}^2 + SE_{bf}{}^2)}$; *p<.05.

-European language. When the regression coefficients were compared to Spanish EL students, Japanese EL students showed differences in all three steps (model fit, model structure, and individual predictor weights), Chinese EL students showed differences in model fit only, and Vietnamese EL students showed differences in individual predictor weight only.

Although the findings have limited generalizability (only one state's data at one grade level), the results provided empirical validity evidence to support the use of ELP assessments to measure EL students' academic language proficiency to perform academic tasks, as indicated by the strong, positive relationship between the ELP and content-area factors. ELP Reading scores were a strong predictor of content-area assessment performance regardless of the different HL groups, indicating that reading skills were crucial for EL students irrespective of their different HL background. This finding suggest that reading intervention and instruction were paramount for all EL students.

Discussion

The MGR analysis is a technique to compare multiple group differences by evaluating the equivalence of the predictive structure of each subgroup's regression model. Unlike traditional statistical hypothesis tests which simply compare the mean scores (e.g., ANOVA), this approach compares three outcomes of multiple regression analysis (model fit, model structure, and individual predictor) using three different types of NHST (Fisher's z, Steiger's z, and Brame et al.'s z). This approach can provide various implications by a) testing regression coefficients for each group and b) comparing regression coefficients across groups. For example, testing one predictor's regression weight for each group can be interpreted as 'one predictor contributes to the regression model for reference group, but not for focal group(s).' On the other hand, comparing one predictor's regression weight across groups can be interpreted as 'one predictor has a larger regression weight in the model for reference group than for focal group(s).' In addition, we propose a practical criterion to evaluate the equivalence of the predictive structure by sequentially testing the equality of regression coefficients. Each statistical hypothesis test directly evaluates the 'equality' of each of the three regression coefficients while the 'equivalence' of the predictive structure is comprehensively evaluated by a combination of three statistical hypothesis tests. This strict criterion allows overcoming the criticism of misusing the NHST (e.g., a statistically non-significant finding does not imply that the groups are comparable; Rusticus & Lovato, 2011).

The foremost advantage of MGR analysis is its simplicity. The procedure is easy to understand and easy to conduct since the group comparison outcomes are based on multiple regression analysis, a well-known statistical technique that analyzes the relationship between one dependent variable and several independent variables. Additionally, the MGR analysis is beneficial due to its flexibility with a small sample size per group compared to the sample size requirement of factor-analytic multi-group analysis (Wilson VanVoorhis & Morgan, 2007). However, the MGR analysis also has its limitations. First, the regression model equation established by the total group should adequately represent each subgroup's model structure; at the very least, identical predictors should be chosen for the reference group model. Suppose all four ELP language domain scores are used to predict the ELA performance from the total EL students model; then, the Spanish group model should include four ELP language domain scores regardless of whether each predictor is statistically significant. If the reference group model's predictors are not equal to the total group model's predictors or the regression coefficients are extremely different between two groups, the interpretation of the group comparison between reference and focal group(s) will be limited and will not be generalizable.

Consequently, selecting a representative and reliable reference group is a fundamental issue in conducting a successful MGR analysis. The sample sizes of subgroups are usually unbalanced in educational research, and the reference group is highly likely the subgroup with the largest sample size. Resampling the large sample groups is suggested in order to avoid the impacts of unequal sample size and variances between samples on statistical power and Type I error rates (Rusticus & Lovato, 2014). In our demonstration case, Spanish EL students were the largest subgroup (43%) of total EL students. A stratified resampling process was thoroughly conducted to create a sampled Spanish group that a) had a similar sample size compared to other focal groups and b) represented the characteristics of both all Spanish EL students and total EL students. The preliminary analysis proved that the regression results from the total EL group were almost identical to those from all Spanish EL students (N=3,165) and those from sampled Spanish EL students (N=306). We strongly recommend meticulously preparing a high-quality reference group for researchers who plan to apply the MGR analysis.

For further research, a simulation study would be insightful in understanding the impact of the interaction between sample sizes and the number of predictors in MGR analysis. In addition, we are interested in extending the current idea of testing the equivalence of regression coefficients into multilevel modeling, which controls the nested data structure (e.g., school-level or district-level).

References

Barnett, R. (2020). Leading with meaning: Why diversity, equity, and inclusion matters in U. S. higher education. *Perspectives in Education*, *38*(2), 20–35.

Brame, R., Paternoster, R., Mazerolle, P., & Piquero, A. (1998). Testing for the equality of maximum-likelihood regression coefficients between two independent equations. *Journal of Quantitative Criminology*, *14*(3), 245–261.

Byrne, B. M. (2012). *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming* (1st ed.). Routledge.

Cardinal, R. N., & Aitken, M. R. F. (2006). *ANOVA for the behavioural sciences researcher*. Mahwah, NJ: Lawrence Erlbaum Associates.

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology, 100*(5), 1261–1293.

Dunn, O. J., & Clark, V. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association, 64*(325), 366–377.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1*, 3–32.

Gay, G. (2018). *Culturally Responsive Teaching: Theory, Research, and Practice*. New York, NY: Teachers College Press.

Hays-Thomas, R. (2016). *Managing Workplace Diversity and Inclusion: A Psychological Perspective*. New York, NY: Routledge.

Hotelling, H. (1940). The selection of variates for use in prediction, with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics, 11*, 271–283.

Jonson, J. L., & Geisinger, K. F. (2022). Introduction: Conceptualizing and Contextualizing Fairness Standards, Issues, and Solutions Across Professional Fields in Education and Psychology. In J. L. Jonson & K. F. Geisinger (Eds.), *Fairness in Educational and Psychological Testing: Examining Theoretical, Research, Practice, and Policy Implications of the 2014 Standards* (pp. 1–10). American Educational Research Association.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*(3), 350–386.

Kline, R. B. (2012). Assumptions in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111–125). The Guilford Press.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling*. New York, NY: The Guildford Press.

Lane, S. (2020). *Test-based accountability systems: The importance of paying attention to consequences* (Research Report No. RR–20–02). Educational Testing Service.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients, *Psychological Bulletin, 111*(1), 172–175.

Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation, 8*(2).

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology, 36*, 859–866.

Russell, M., Szendey, O., & Li, Z. (2022). An intersectional approach to DIF: Comparing outcomes across methods. *Educational Assessment, 27*(2), 115–135.

Rusticus, S., & Lovato, C. (2011). Applying tests of equivalence for multiple group comparisons: Demonstration of the confidence interval approach. *Practical Assessment, Research, & Evaluation*, *16*(7).

Rusticus, S., & Lovato, C. (2014). Impact of sample size and variability on the Power and Type I error rates of equivalence tests: A simulation study. *Practical Assessment, Research & Evaluation. 19*(11).

Sorböm, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229–239.

Steiger, J. H. (1980) Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminancy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis significance tests. *Psychological Methods, 6*, 371–386.

Wilson VanVoorhis, C. R., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology, 3*(2), 43–50.

Wolf, M. K., Bailey, A. L., Ballard, L., Wang, Y., & Pogossian, A. (2022). Unpacking the language demands in academic content and English language proficiency standards for English learners. *International Multilingual Research Journal, 17*(1), 68–85.

Citation:
Yoo, H., Wolf, M. K., & Ballard, L. D. (2023). Evaluating the equality of regression coefficients for multiple group comparisons: A case of English learner subgroups by home languages. *Practical Assessment, Research, & Evaluation*, 28(5). Available online: https://scholarworks.umass.edu/pare/vol28/iss1/5/

Corresponding Author:
Hanwook Yoo
Educational Testing Service

Email: hyoo [at] ets.org

Appendix A. EL Students' ELP and Content Assessment Performances

Figure A1. EL students' ELP Assessment Performance by Home Language

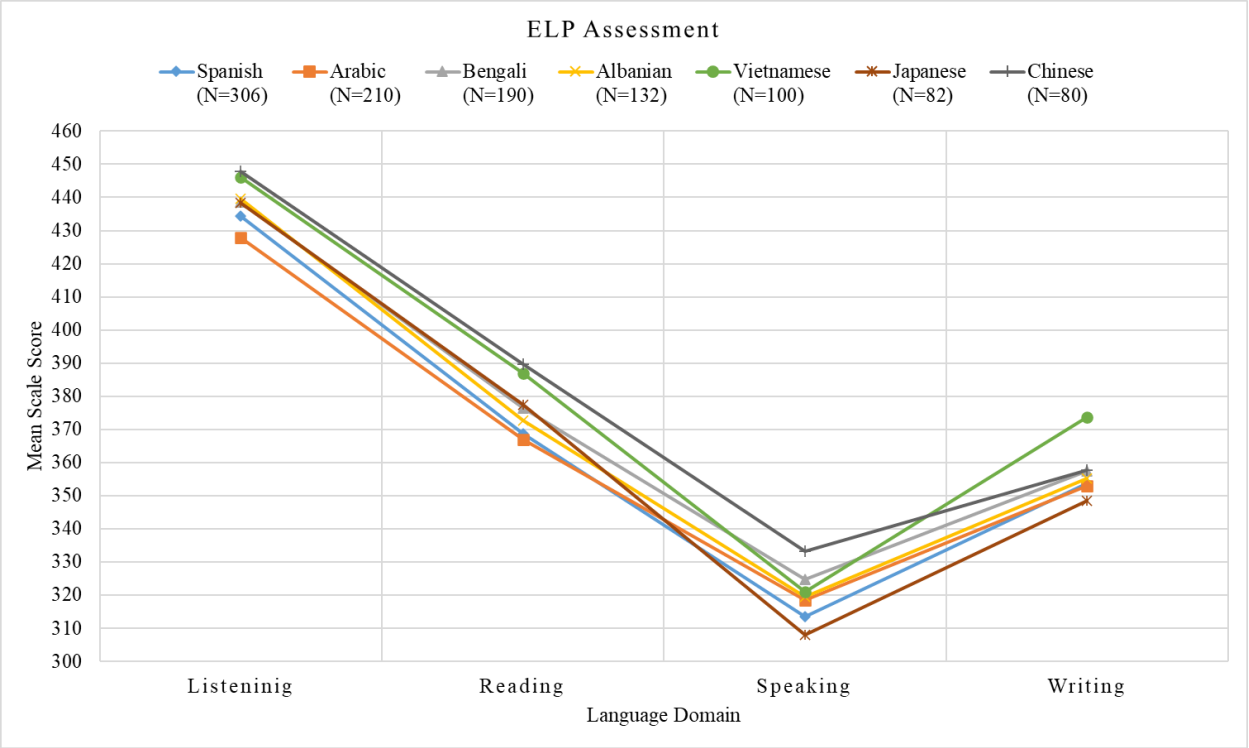
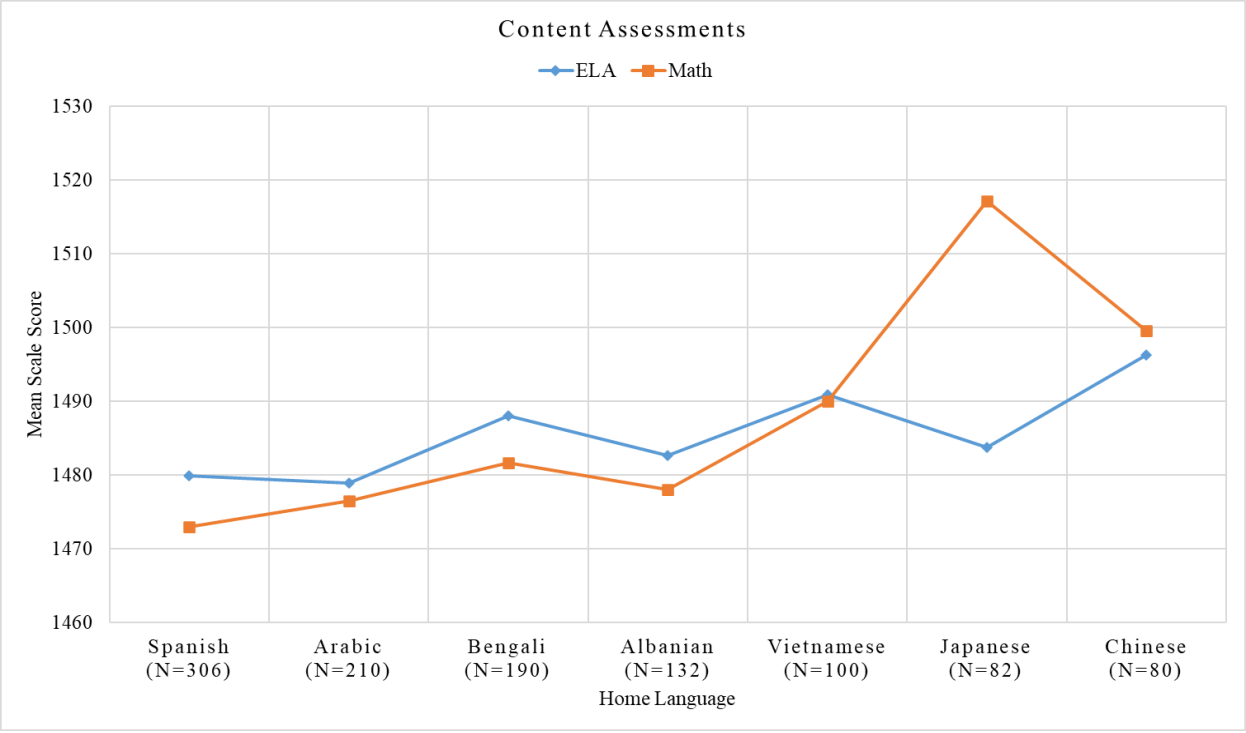


Figure A2. EL students' Content Assessment Performance by Home Language



Appendix B. Comparisons of Individual Predictor Weights (Predicted ELA Model)

Table B1. Individual Predictor Comparison (ELA; Arabic HL group)

Predictor	Spanish		Arabic		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.07	0.03	0.06	0.03	0.04	0.271	0.786
Reading	0.42	0.04	0.36	0.05	0.06	1.064	0.287
Speaking	0.02	0.02	0.03	0.03	0.04	-0.503	0.615
Writing	0.01	0.03	0.01	0.04	0.05	-0.040	0.968

Table B2. Individual Predictor Comparison (ELA; Bengali HL group)

Predictor	Spanish		Bengali		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.07	0.03	0.14	0.04	0.05	-1.523	0.128
Reading	0.42	0.04	0.42	0.06	0.07	0.023	0.982
Speaking	0.02	0.02	-0.04	0.03	0.04	1.400	0.162
Writing	0.01	0.03	-0.04	0.05	0.06	0.892	0.372

Table B3. Individual Predictor Comparison (ELA; Albanian HL group)

Predictor	Spanish		Albanian		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.07	0.03	0.04	0.04	0.04	0.691	0.490
Reading	0.42	0.04	0.41	0.05	0.06	0.240	0.810
Speaking	0.02	0.02	0.00	0.03	0.04	0.409	0.683
Writing	0.01	0.03	0.03	0.05	0.05	-0.388	0.698

Table B4. Individual Predictor Comparison (ELA; Vietnamese HL group)

Predictor	Spanish		Vietnamese		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.07	0.03	0.16	0.05	0.05	-1.764	0.078
Reading	0.42	0.04	0.39	0.07	0.08	0.365	0.715
Speaking	0.02	0.02	0.03	0.04	0.04	-0.281	0.779
Writing	0.01	0.03	-0.02	0.07	0.07	0.468	0.640

Table B5. Individual Predictor Comparison (ELA; Japanese HL group)

Predictor	Spanish		Japanese		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.07	0.03	0.16	0.05	0.06	-1.709	0.087
Reading	0.42	0.04	0.20	0.07	0.08	2.797	0.005**
Speaking	0.02	0.02	0.12	0.05	0.05	-2.161	0.031*
Writing	0.01	0.03	0.04	0.07	0.08	-0.415	0.678

Table B6. Individual Predictor Comparison (ELA; Chinese HL group)

Predictor	Spanish		Chinese		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.07	0.03	0.09	0.05	0.05	-0.338	0.735
Reading	0.42	0.04	0.45	0.06	0.07	-0.427	0.669
Speaking	0.02	0.02	0.00	0.04	0.04	0.361	0.718
Writing	0.01	0.03	0.04	0.07	0.07	-0.381	0.703

Appendix C. Comparisons of Individual Predictor Weights (Predicted Math Model)

Table C1. Individual Predictor Comparison (Mathematics; Arabic HL group)

Predictor	Spanish		Arabic		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.08	0.03	0.12	0.04	0.05	-0.819	0.413
Reading	0.34	0.04	0.31	0.05	0.07	0.406	0.685
Speaking	0.04	0.03	0.03	0.03	0.04	0.166	0.868
Writing	-0.03	0.03	-0.02	0.05	0.06	-0.172	0.863

Table C2. Individual Predictor Comparison (Mathematics; Bengali HL group)

Predictor	Spanish		Bengali		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.08	0.03	0.16	0.04	0.05	-1.607	0.108
Reading	0.34	0.04	0.30	0.06	0.07	0.522	0.602
Speaking	0.04	0.03	0.04	0.03	0.04	-0.039	0.969
Writing	-0.03	0.03	-0.06	0.05	0.06	0.498	0.618

Table C3. Individual Predictor Comparison (Mathematics; Albanian HL group)

Predictor	Spanish		Albanian		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.08	0.03	0.07	0.04	0.05	0.241	0.810
Reading	0.34	0.04	0.39	0.06	0.08	-0.614	0.539
Speaking	0.04	0.03	0.06	0.04	0.05	-0.428	0.669
Writing	-0.03	0.03	-0.04	0.06	0.06	0.154	0.878

Table C4. Individual Predictor Comparison (Mathematics; Vietnamese HL group)

Predictor	Spanish		Vietnamese		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.08	0.03	0.19	0.05	0.06	-1.835	0.067
Reading	0.34	0.04	0.16	0.07	0.08	2.158	0.031*
Speaking	0.04	0.03	0.04	0.04	0.05	-0.114	0.909
Writing	-0.03	0.03	0.06	0.07	0.08	-1.122	0.262

Table C5. Individual Predictor Comparison (Mathematics; Japanese HL group)

Predictor	Spanish		Japanese		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.08	0.03	0.05	0.05	0.06	0.549	0.583
Reading	0.34	0.04	0.24	0.07	0.08	1.193	0.233
Speaking	0.04	0.03	-0.05	0.05	0.05	1.642	0.101
Writing	-0.03	0.03	0.05	0.07	0.08	-1.063	0.288

Table C6. Individual Predictor Comparison (Mathematics; Chinese HL group)

Predictor	Spanish		Chinese		SE(B-diff)	Z	p
	B	SE	B	SE			
Listening	0.08	0.03	0.06	0.06	0.07	0.288	0.773
Reading	0.34	0.04	0.40	0.08	0.09	-0.644	0.520
Speaking	0.04	0.03	-0.05	0.05	0.06	1.459	0.145
Writing	-0.03	0.03	-0.07	0.09	0.10	0.414	0.679