Journal of Educational Psychology
2008, Vol. 100, No. 2, 322–342

0022-0663/08/$12.00 DOI: 10.1037/0022-0663.100.2.322

Native Language Proficiency, English Literacy, Academic Achievement, and Occupational Attainment in Limited-English-Proficient Students: A Latent Growth Modeling Perspective

R. Sergio Guglielmi
Lake Forest College

The hypothesis that native language (L1) proficiency promotes English acquisition and overall academic achievement, a key theoretical assumption underlying bilingual education, was tested using latent growth modeling of data from 899 limited-English-proficient (LEP) eighth graders who were followed for 12 years in the National Education Longitudinal Study (NELS:88/2000). A model in which L1 proficiency predicted English (L2) reading ability, which in turn predicted high school achievement and distal educational/occupational attainment, fit the data well for the full LEP sample and a Hispanic subsample. In Hispanics, the model explained 24.1%, 7.4%, 29.4%, and 46.3% of the variance in initial English reading level, English reading growth, high school achievement, and post–high school attainment, respectively. Model fit for an Asian subsample, however, was poor. Possible reasons for lack of group invariance include cultural differences in construct conceptualization, greater linguistic and cultural heterogeneity within the Asian subgroup, and cross-language transfer difficulties when L1 and L2 lack a shared alphabetic structure. At least for Hispanic LEP students, this study's results establish the theoretical foundation for exploring the effectiveness of specific educational interventions.

Keywords: bilingual education, English language learner, latent growth modeling, multitrait–multimethod, literacy

The U.S. Department of Education (Kindler, 2002) estimates that in 2000–2001 more than 4.5 million public school students (preK–12) in the United States were identified as limited-English-proficient (LEP),[1] and by the year 2030, this number is projected to grow to 40% of the school-age population (Thomas & Collier, 2002). How our schools should respond to the needs of the rapidly expanding LEP population has been the object of a vigorous national debate.

In *Lau v. Nichols* (1974), the U.S. Supreme Court recognized that LEP students would be locked out of the educational system unless schools developed instructional programs that would give them access to a meaningful education despite the language barrier. Although the Court prescribed action, it gave local school districts the power to decide which particular educational programs to institute. Since then, legislators, educators, academicians, policymakers, and advocacy groups have argued the relative merits of various approaches to educating the growing LEP population. The controversy over the effectiveness of bilingual education has reached especially harsh tones. Fundamentally, the debate centers on the value of two different pedagogical models. One approach is full English immersion. In these cases, the language barrier problem is often addressed by placing LEP students in English-as-a-second-language (ESL) classes, where they receive instruction in the English language skills necessary to operate in a mainstream classroom. On the other side of the debate, proponents of bilingual education advocate academic instruction in both the students' first language (L1) and in English, the second language (L2), with the amount of time spent in L1 instruction decreasing progressively over the course of a few years (early exit programs) or several years (late exit programs).

In the last several years, bilingual education has come under attack, and in some cases (e.g., the passage of Proposition 227 in California and of Proposition 203 in Arizona), its continued existence has been threatened. Unfortunately, judgments about the effectiveness of bilingual education (for or against) are often driven more by sociopolitical motivations than by an objective evaluation of the scientific evidence. A substantial literature on the

Editor's Note. Linda Baker served as the action editor for this article.—KRH

This work was made possible, in part, by a grant from Carnegie Corporation of New York. The statements made and views expressed in this article are solely the responsibility of the author. I am grateful to Lake Forest College for a sabbatical leave that provided the time necessary to complete this project; Nancy Brekke, Michael Nettles, and Sean Reardon for their careful reading of a draft and their constructive feedback; Laura Stapleton for useful suggestions about the analysis of multilevel longitudinal data in a complex sample framework; and Amber Trujillo for her insightful perspective on the challenges and controversies concerning the education of language minorities.

Correspondence concerning this article should be addressed to Sergio Guglielmi, Department of Psychology, Hotchkiss Hall, Box E4, Lake Forest College, 555 North Sheridan Road, Lake Forest, IL 60045. E-mail: guglielm@lakeforest.edu

[1] Although the term English language learner (ELL) has become preferable in the last few years, I will continue to refer to these students as LEP in order to maintain consistency with the terminology most often adopted in legislative documents, in government publications, and in the data set used in the present investigation.

Table 1
Multitrait–Multimethod Correlation Matrix for Two Assessment Methods and Two Traits Measured at Time 1 and at Time 2

	Self-report				Objective assessment			
Method/measure	1	2	3	4	5	6	7	8
Self-report								
1. L2 proficiency at T1	—							
2. Non-English grades at T1	−.051	—						
3. L2 proficiency at T2	.542	.029	—					
4. Non-English grades at T2	−.038	.470	−.064	—				
Objective assessment								
5. IRT reading scores at T1	.297	.328	.324	.247	—			
6. Non-English GPA at T1	.085	.379	.134	.499	.413	—		
7. IRT reading scores at T2	.261	.307	.315	.273	.773	.451	—	
8. Non-English GPA at T2	−.026	.379	.082	.578	.311	.611	.392	—

Note. Because of the large sample size ($N = 1{,}804$), correlations as small as .047 are significant at $p < .05$. T1 = Time 1 data collection; T2 = Time 2 data collection; IRT = item response theory, GPA = grade point average.

Construct Validity: A Multitrait–Multimethod Model

A second important issue concerns the construct validity of the L1 and L2 proficiency measures that are based on self-report and are therefore vulnerable to selective distortion by self-presentational biases and other response sets. Students' self-rated proficiency measures reflect three sources of variance: (a) trait variance, which is systematic variance, independent of assessment method, that represents true differences in ability level; (b) method variance, which is systematic variance specifically associated with the use of self-report (e.g., the tendency to be influenced by social desirability concerns); and (c) error variance, which includes both residual systematic variance and measurement error (e.g., misreading or misunderstanding a question). In order to determine whether self-ratings of language proficiency reflect true variations in ability, one must disentangle the contributions of trait effects, method effects, and error effects. Moreover, if L1 or L2 proficiency were found to predict favorable academic and occupational outcomes, it would be important to rule out the possibility that those associations were simply the byproduct of a general intellectual ability factor.

The multitrait–multimethod (MTMM) approach pioneered by Campbell and Fiske (1959) can be used to decompose the total variance and to assess the convergent and discriminant validity of psychological measures. This design requires that two or more traits be assessed with two or more methods. Regrettably, the NELS data set provides no measure of L1 proficiency other than self-report. More objective indices of proficiency, against which one could validate students' self-ratings, however, are available for L2. Table 1 shows a MTMM matrix that includes two methods (self-report and objective measurement) and two traits (L2 proficiency and achievement in academic subjects other than English) assessed at two points in time (Grades 8 and 10).[4] The objective indicators of the two traits are based on transcripts and on standardized test scores. The non-English achievement indicators, included to test the discriminant validity of the L2 proficiency measure, represent self-reported and transcript-based grades in math, science, and social studies courses.

Longitudinal confirmatory factor analysis (CFA) of the MTMM matrix was used to evaluate the construct validity and temporal stability of self-reported L2 proficiency. The CFA model (see Figure 2) includes two trait factors across two waves, two method factors, and eight observed variables. For each wave, one self-report variable and one objective assessment variable were specified to load on each trait.[5] The manifest variables that at each wave were hypothesized to load on the L2 proficiency trait factor were the IRT-estimated reading scores and a self-reported proficiency measure computed by averaging students' self-rated ability to understand, speak, read, and write English.[6] The two indicators that were specified to load at each wave on the Non-English Achievement trait factor were (a) non-English grades, which were computed by averaging students' self-reported grades in math, science, and social studies, and (b) non-English GPA, which was

[4] Although L2 proficiency ratings are available also at Time 3 (second follow-up questionnaire), a self-report measure of grades in various academic subjects was not included at that time. Thus, only the first two waves of data could be used to test the model.

[5] There is considerable debate on the advantages and disadvantages of various CFA parameterizations of MTMM matrices (see Brown, 2006, for a recent review). The correlated trait-correlated method (CT-CM) approach was used in the present research for the theoretical and substantive reasons articulated by Conway, Lievens, Scullen, and Lance (2004) and by Lance, Noble, and Scullen (2002). This parameterization, however, is notorious for returning ill-defined solutions (Kenny & Kashy, 1992). The problem is often remedied with the use of large samples (e.g., Lance et al., 2002). Since the data to be modeled no longer required participation in all waves of the study, the CFA–MTMM model was tested on 1,804 participants who met the inclusion criteria described in Figure 1, participated in the first and second follow-ups, and had transcript information. The data were weighted with "f2trp1wt," the appropriate NELS weight in panel analyses in which survey and cognitive test data are combined with transcript data.

[6] In an alternative model, the mean of the four language skills was replaced with only the reading self-rated skill that was specified to load, together with the IRT reading measure, on a L2 reading proficiency trait factor. Estimation of this model produced a "not positive definite matrix" solution, probably as a result of a very high correlation between Time 1 and Time 2 reading proficiency factors ($r = .993$). The fit of this model, the parameter estimates, and the standard errors, however, were substantively interchangeable with those of the model shown in Figure 2.

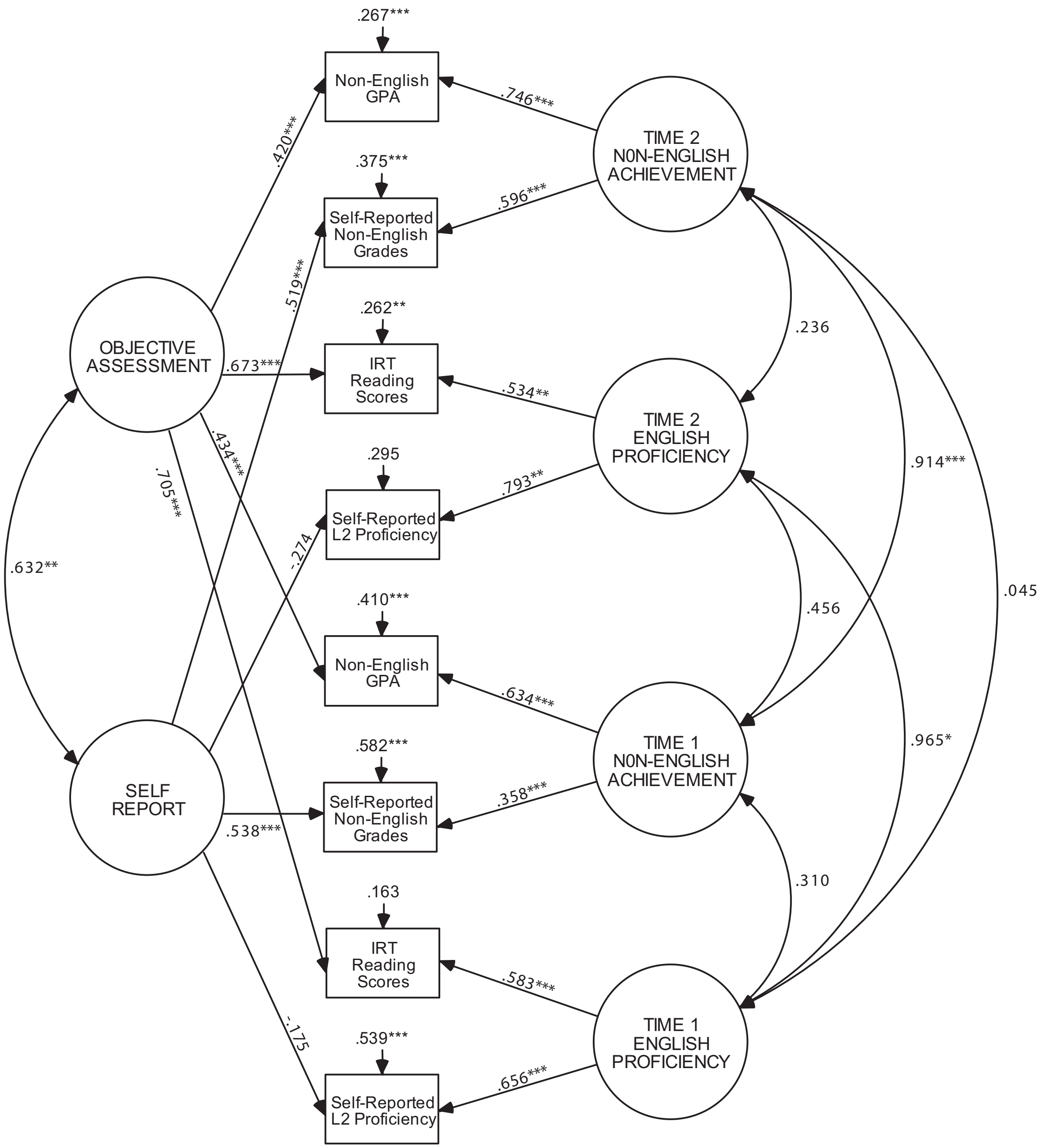

Figure 2. Multitrait–multimethod confirmatory factor analysis model of two correlated traits and two correlated methods across two measurement waves. Completely standardized robust maximum likelihood parameter estimates. The residual variance components (error variances) indicate the amount of unexplained variance. Thus, for each observed variable, $R^2 = (1 - \text{error variance})$. GPA = grade point average; IRT = item response theory; L2 = English. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

the transcript-based average of students' GPA in math, science, and social studies.[7]

The fit of the CFA–MTMM model displayed in Figure 2 was excellent, $\chi^2(5, N = 1{,}804) = 10.236$, CFI = .994, RMSEA = .024, SRMR = .015 (see Model 1 in Table 2, which summarizes the fit of all the models tested in the present investigation). For both waves, all measures loaded significantly ($p < .001$) on their respective trait factors, even after shared method variance was controlled. This indicated good levels of convergent validity. Discriminant validity of the measures was demonstrated by the non-

[7] The non-English GPA variable for Time 2 is an average of standardized course grades obtained in 10th grade. The NELS transcript file, however, includes eighth-grade GPA for only a handful of students; thus ninth-grade courses were used to derive the Time 1 non-English GPA variable.

Table 2
Fit Indices for the Models Tested

Model	Comparison model	χ^2	*df*	CFI	RMSEA	SRMR	Δ_{χ^2}	Δ_{df}	Δ_{CFI}
Single group analyses									
1. MTMM	—	10.236	5	.994	.024	.015	—	—	—
2. Growth	—	4.290*	1	.997	.061	.015	—	—	—
3. CFA	—	190.213*	82	.961	.038	.043	—	—	—
4. Full conditional LGM	—	343.074*	150	.952	.038	.050	—	—	—
5. Full LGM (Hispanic sample)	—	355.490*	150	.934	.055	.064	—	—	—
6. Model 5, but paths from L1 proficiency to growth factors fixed at 0	5	391.113*	152	.923	.059	.092	79.052*	2	.011
7. Model 5, but paths from L2 proficiency to growth factors fixed at 0	5	361.884*	152	.932	.055	.070	6.248*	2	.002
8. Full LGM (Asian sample)	—	641.974*	150	.778	.114	.098	—	—	—
9. Model 8, but paths from L1 proficiency to growth factors fixed at 0	8				Improper solution (negative Δ_{χ^2})				
10. Model 8, but paths from L2 proficiency to growth factors fixed at 0	8	643.184*	152	.778	.113	.103	5.218	2	.000
Multigroup invariance tests									
11. Unconstrained growth model—baseline	—	2.632	2	.998	.030	.016	—	—	—
12. Model 11, but growth factor means invariant	11	7.027	4	.990	.046	.073	4.018	2	.008
13. Model 12, but growth factor variances & covariances invariant	12	10.538	7	.988	.038	.084	3.529	3	.002
14. Unconstrained CFA model—baseline	—	580.206*	164	.889	.085	.070	—	—	—

Note. χ^2 = Yuan–Bentler corrected χ^2; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; MTMM = multitrait–multimethod approach; CFA = confirmatory factor analysis; LGM = latent growth modeling; L1 = native language; L2 = English.
*$p < .05$.

significant correlations between different trait factors, both within and across measurement occasions. At the same time, the very high correlations for the same trait factor across time indicated excellent temporal stability of the constructs. Table 1 shows that heterotrait–monomethod correlations were generally higher in the objective assessment block than in the self-report block, indicating greater shared method variance for the objective measures than for the self-report measures. The squared factor loadings of each measure on its respective trait and method factors made it possible to partition the total variance into trait, method, and error components. Again, large method effects were evident for the objective measures, all of which loaded significantly ($p < .001$) on their method factor. Method effects were smaller but also significant for the self-report non-English grade measures and were nonsignificant for the self-report L2 proficiency indicators. Noteworthy was the large amount of trait variance relative to method variance for both waves, particularly for the self-report measures. Overall, averaging variance components across all measures and across the two measurement occasions indicated that 39.1% of the variance in the measures was explained by trait variance, 24.7% was explained by method variance, and the remaining 36.2% was due to error.

Taken together, these findings are reassuring with regard to the reliability, stability, and construct validity of the L2 proficiency self-ratings. The L2 proficiency measure loaded strongly on the same trait factor on which IRT-estimated English reading scores loaded significantly. In addition, the nonsignificant correlations between L2 proficiency and academic achievement in non-English subjects suggest that L2 skills were not byproducts of a general cognitive ability factor. It seems reasonable to extend to the L1 proficiency measure the favorable conclusions reached about the psychometric characteristics of the L2 proficiency variable, particularly considering that criterion-related validity coefficients have repeatedly been found to be higher for L1 than for L2 self-assessment (e.g., Delgado et al., 1999; Hakuta & D'Andrea, 1992).

Latent Growth Curve Analyses

The key hypothesis under examination in this study was that L1 proficiency would predict the development of L2 reading skills in LEP students, and this, in turn, would be associated with successful academic and occupational outcomes. IRT English reading scores from three points in time—8th, 10th, and 12th grades—were available. Thus, the first step was to establish whether this sample of LEP students showed evidence of growth in reading scores and whether there were substantial individual differences in growth. Furthermore, prior to evaluating the structural component of the model, the validity of the measurement model needed to be assessed. Assuming confirmation that the manifest indicators adequately measured their respective latent factors and assuming a reasonable level of interindividual variation in initial reading scores and rate of change, one could then evaluate the hypothesized relations among latent constructs.

The longitudinal structure of the data makes these research questions ideally suited for latent growth modeling (LGM) analyses, which become possible when at least three measurement occasions are available for the repeated measure variable. Although multistep modeling strategies have been the object of some controversy (e.g., see *Structural Equation Modeling* [2002], *7,* [1], the analytic approach adopted in the present research followed