Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 25

A Multi-group Generalizability Analysis of a Large-scale Reading Comprehension Test.

Dongmei Li
Robert L. Brennan†

August 2007

† Robert L.Brennan is E.F.Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist, College of Education, University of Iowa, Iowa City, IA 52242 (email: robert-brennan@uiowa.edu). Dongmei Li is a research assistant in Iowa Testing Programs (email: dongmei-li@uiowa.edu)

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Abstract

Though it is common for a large-scale standardized test to report just one number as its reliability index, the one index may no longer hold when different sources of error are taken into account or when it is used for subgroups with very different characteristics. For a reading comprehension test, it is usually the case that several passages are used with various numbers of items in each, with each of the items targeting one of the sub-skills or cognitive processes involved in reading as specified by the test developers. Therefore, items, passages, and cognitive processes are all facets to be considered for reliability estimation. At the same time, since the same reading comprehension test is often taken by students who are English Language Learners (ELLs) as well as by those who are native English speakers, it is suspected that reliability estimates won't be the same for these two groups. By conducting a series of generalizability analyses of a reading comprehension test for both groups, this study demonstrated the amount of discrepancy in coefficients and error variances when different facets are taken into account, and the differential contribution of these facets to measurement error for ELLs and native English speakers.

A Multi-group Generalizability Analysis of a Large-scale Reading Comprehension Test

Facets of a Reading Comprehension Test

Some researchers view reading comprehension as a unitary, holistic and indivisible skill (e.g. Alderson, 1990a, b; Rost, 1993). Others, however, view it as composed of various sub-skills(e.g. Bloom, 1956; Davis, 1968). In the assessment of reading comprehension, the latter view is often adopted by the test developers so that items are developed with each specifically targeting one of the sub-skills or cognitive processes involved in reading. In a test of reading comprehension, therefore, it is usually the case that several passages are used with various numbers of items in each, with each of the items targeting one of the sub-skills or cognitive processes involved in reading as specified by the test developers. When considered from the perspective of generalizability theory, the measurement of reading proficiency using such typical tests can be viewed as involving three facets which can all influence the generalizability of scores, that is, passages, processes, and items that are nested within the cross classification of passages and processes.

In the literature, many studies have investigated the passage effect on estimation of reliability or generalizability (e.g. Wan & Brennan, 2005; Lee & Frisbie, 1999). However, the effect of cognitive processes on reliability has not been investigated so frequently. One of the reasons might be that the process dimension is not directly observable. Even though items are often classified into various cognitive categories, the actual cognitive processes used by the examinees are usually not known.

Nevertheless, cognitive processes are of great practical importance. First, they can provide diagnostic information on students' cognitive strengths and weaknesses that testing programs would like to include in their score reports. Second, this may be an important dimension for individual differences and group differences on reading proficiency. For example, cognitive processes were found to account for a large proportion of the variances of differential item functioning indices of a reading test between English Language Learners and those who are not (Li & Kolen, 2005). Given the importance of cognitive processes in both test development and test score interpretation, it is desirable that this facet be evaluated with respect to reading score generalizability.

Group Differences in the Test Population

When provided with the value of a reliability or generalizability index for test scores, it is often assumed that the test is equally reliable for all subgroups in the test population. However, this assumption may not hold when systematic differences exist among groups of examinees, such as groups based on ethnicity, gender, and/or geographic regions. Another pair of commonly contrasted groups that have gained more and more attention in recent years is the group of English language learners and the group of native English speakers.

English language learners, or simply ELL, is a term used in the K-12 system referring to students whose native language is not English and who are still in the process of learning and acquiring the English language. In the United States, the number of ELL students has been increasing rapidly over the last few decades with the flow of immigration, and recently reached 10% of the public school enrollment (Kinder, 2002).

With the policy requirement of inclusion of all students including ELL students in standardized assessment for the purpose of school accountability, more and more ELL students need to take the same tests as non-ELL students, and some important educational decisions will be based on test scores.

However, the differences between ELL students and non-ELL students are a source of concern for the validity of ELL assessment. For example, these ELL students not only differ from their native English speaking peers in terms of English language proficiency, they also differ in educational and cultural background. Studies have been conducted to investigate the group differences between ELL students and non-ELL students in terms of test performances (Abedi, Coutney, & Leon, 2003; Abedi, Lord, & Hofstetter,1998; Butler & Stevens, 2001; Mao& Li, 2003), differential item functioning (Li & Kolen, 2005), or construct equivalence (Wan, Li, & Dunbar, 2006). Some researchers proposed the use of generalizability theory to examine differences between groups (e.g. Abedi, 2002; Solano-Flores & Trumbull, 2003), yet few if any studies have actually conducted such analyses.

Purpose of the Study

This study has two purposes. One is to investigate the differences in generalizability indices when different facets of a reading comprehension test are taken into account. The other is to demonstrate the examination of group differences using generalizability analyses by comparing different sources of error variances for ELL students and non-ELL students.

Method

Data and instrument

The data were responses from 500 ELL students and 500 non-ELL students on ITBS Reading Comprehension (Level 13). These data were random samples of the population of ELL students and the population of non-ELL students from an administration of the test in 2003.

The reading comprehension test consists of 8 passages and 48 items, each of which is classified into one of the following cognitive process categories: *factual understanding* (FACT), *inference & interpretation* (INFER), and *analysis & generalization* (GENER). The number of items in each passage ranges from 4-7, and in each process category 14-17. The distribution of items in each passage crossed with cognitive process category is listed in Table 1. There are 0-4 items in each cell.

Generalizability Analyses

Generalizability theory provides a framework to conceptualize and disentangle multiple sources of error. For a reading comprehension test, with persons (***p***) as the object of measurement, three facets contribute to the person score variability, i.e., passages (***h***), items (***i***), and cognitive processes (***c***). It is usually the case that for each passage, there are items that are intended for each one of the cognitive processes, and all persons are administered the same sets of items. Whereas the passages are often considered random samples from a large pool of appropriate passages, the cognitive processes are usually considered to be fixed as stated in the test specifications.

Given the above conceptualization of the universe of admissible observations, a multivariate random facet $p^{\bullet} \times (i^{\circ} : h^{\bullet})$ G study design would be most appropriate

(Brennan, 2001a). The notation used in this study follows Brennan (2001a). The superscript filled circle $\bullet$ indicates that the facet is crossed with the fixed multivariate variables, and the superscript empty circle $\circ$ indicates that the facet is nested within the fixed multivariate variable. No superscripts are used for univariate designs.

However, one or more of the facets are often ignored in practice, leading to biased estimates of the generalizability of a reading comprehension test. This study considered four separate generaliabiliity analyses: (1) a univariate $p \times i$ analysis that ignored both the passage and process facets, (2) a multivariate $p^{\bullet} \times i^{\circ}$ analysis that took process into consideration as a fixed facet but ignored the passage facet, (3) a univariate $p \times (i : h)$ analysis that took into account the passage facet but ignored the process facet, and (4) a multivariate $p \times (i^{\circ} : h^{\bullet})$ analysis that took into account both the passage and process facets. A G study and a default D study that had the same design and the same sample size for each facet were conducted for all the above four analyses.

Additional D studies were conducted to answer various questions. First, for further examination of the processes, different sample sizes and different *a prior* weights were used to investigate how the generalizability of subscores and composite scores would improve with longer tests and what difference it would make when the processes were weighted equally or differentially. In addition, the variability of the profiles of process scores and conditional standard errors of measurement were also examined. Second, to further examine the passage effect, different combinations of passages and items were explored.

All the analyses were conducted using mGENOVA (Brennan, 2001b).

Passage effect and process effect

Theoretically, the univariate $p \times i$ analysis is the least appropriate for a reading comprehension test intended to assess multiple cognitive processes, yet it is actually the most commonly used in practice due to simplicity. Testing programs routinely report reliability indices based only on the item facet, and a large amount of research in the literature does the same. It has been demonstrated repeatedly in the previously cited studies that reliability is overestimated when using the $p \times i$ design rather than the more appropriate univariate $p \times (i : h)$ design. This study examines the degree of overestimation again in the context of this particular reading comprehension test by comparing results from designs (1) and (3) as well as those from (2) and (4).

With respect to the cognitive processes facet, it is expected that incorporating this fixed facet may boost the estimate of reliability a little bit due to the contribution of some process related variance to the universe score variance, yet few studies have empirically investigated this. This study tries to determine the contribution of cognitive processes to the estimation of the generalizability index by comparing results from designs (1) and (2) as well as (3) and (4).

Group comparison

These analyses were done for the ELL group and the non-ELL group separately and group differences were examined at the same time. Additional D studies were conducted for the ELL group using the multivariate $p^{\bullet} \times (i^{\circ} : h^{\bullet})$ design to investigate changes in sample sizes necessary to achieve generalizability comparable to the non-ELL group .

Results

To give a general idea about the group differences in the reading comprehension test, the performances of ELL students and non-ELL students are plotted in Figure 1 in terms of mean scores on each cognitive process, and in Figure 2 in terms of mean scores on each passage. Next, results for each of the four generalizability study designs are reported respectively.

Univariate $p \times i$ study results

The G and D study results of the $p \times i$ design for both groups are reported in Table 2. The generalizability coefficient for the non-ELL group is 0.917, and for the ELL group .852. The lower coefficient for the ELL group results from the smaller universe score variance but larger relative error variance of this group compared to the non-ELL group. The fitted conditional relative SEM (CSEM) for both groups are plotted in Figure 3. Little difference is observed between groups regarding CSEM.

Multivariate $p^{\bullet} \times i^{\circ}$ study results

Figure 1 plots the mean scores for each group on the three cognitive processes. The multivariate $p^{\bullet} \times i^{\circ}$ G study results are reported in Table 3. On the left portion of the table are statistics used to estimate the variance and covariance components, and on the right portion of the table are the estimated variance and covariance components and disattenuated correlations. Again, the persons in the non-ELL groups are more variable than those in the ELL group. From the upper diagonal of the p matrices, it can be seen that though the observed correlations between processes for the ELL group are all smaller than for the non-ELL group, the disattenuated correlations for both groups are all extremely high. The lowest disattenuated correlation is between INFER and GENER for

the non-ELL group which is .943. All other correlations are very close to or slightly over 1.0.

The D study results for both groups using the same sample sizes and design as the G study are reported in Table 4. A larger gap in the generalizability and dependability coefficients is observed between groups for the three process variables. While the generalizability coefficients for the non-ELL group for individual variables are all in the upper .70's, they are all in the .60's for the ELL group, with the smallest gap between groups being .102 for FACT, and the largest being .153 for GENER. The differences in the dependability coefficients between groups have a similar magnitude and pattern. The coefficients for the sample-size weighted composite scores are very close to those obtained using the $p \times i$ design for both groups.

The contributions of each variable to the overall universe score variance and error variances are reported at the lower portion of Table 4. For both groups, GENER contributes least to the universe and error variances, which is partly (if not largely) attributable to a smaller number of items in this process category than in the other two categories.

Tables 5 and 6 report results for two more D studies for the non-ELL group and ELL group, respectively. One D study investigated the use of the same sample sizes as the G study but equal a priori weights (w-weights) for the three processes, and the other one investigated the use of more items given equal weights. It is clear that when more items are used for each process, the generalizability and dependability coefficients for both groups increase. However, even with a test that is one half longer than the test of the non-ELL students, the coefficients for the ELL students would still be slightly lower for

each individual variable as well as for the composite score. It is also interesting to note that given the original test, if the three processes were weighted equally, GENER would contribute more to the universe score variance and error variances for both groups than the other two variables.

Figures 4 plots the conditional relative SEM for both groups. Again, little difference can be observed between groups.

Univariate $p \times (i:h)$ *study results*

The univariate $p \times (i:h)$ G study and the default D study results for the non-ELL group are reported in Table 7, and those for the ELL group are in Table 8. Results of the G studies show that for both groups, the variance related to passage is several times smaller than that related to items within passage. However, the passages have a larger variance for the ELL group than for the non-ELL group. The Default D study results for the two groups again show lower generalizability of scores for the ELL group than for the non-ELL group.

Table 7 and 8 also show results of a few additional D studies. One set of D studies conducted for both groups is the exploration of different combinations of passage numbers and item numbers while maintaining the total number of items in the present test. Results for both groups are consistent in that increasing the number of passages can raise the generalizability of the test faster than increasing the numbers of items within passages. Another set of D studies are reported for the ELL group, indicating the number of additional items or passages that are needed for this group in order to approach equal generalizability as the non-ELL group.

The conditional relative standard errors of measurement for both groups are plotted in Figure 5. More difference is observed between the groups at the lower end of the score scale, where the CSEM is smaller for the ELL group than for the non-ELL group[1].

Multivariate $p^{\bullet} \times (i^{\circ} : h^{\bullet})$ *study results*

Table 9 reports the G study results for both groups. Consistent with the multivariate $p^{\bullet} \times i^{\circ}$ study results, the disattenuated correlations between cognitive processes are close to perfect. For all cognitive processes, variances related to passages are again much smaller than variances related to items nested within passages across groups. However, when comparing the relative magnitude of the variance components of different effects between groups, the patterns may be different across cognitive processes. For example, while the non-ELL group has a larger variance for the passages regarding FACT, the ELL group has a larger variance regarding GENER.

The default D study results are reported in Table 10. The relative patterns among variance components and generalizabiility and dependability indices are consistent with those of the multivariate $p^{\bullet} \times i^{\circ}$ design. Additional D studies were also conducted for this design including an exploration of combinations of numbers of passages and items, and increased test lengths for the ELLs. Results are not reported here because they are very similar with earlier reported results for simpler designs.

The conditional relative standard errors of measurement for both groups are plotted in Figure 6. The pattern of difference between the two groups is very similar with that of the univariate $p \times (i : h)$ design as shown in Figure 5.

[1] The smaller CSEM for the ELL group does not contradict the lower generalizability for this group because CSEM is the within person variability which does not account for the universe score variance.

Summary of results for comparison across designs and between groups

In order to facilitate comparison of results across designs and between groups, some results concerning the reading comprehension test itself (i.e. default sample sizes) are reorganized and re-reported in Table 11, and Tables 12-15 are created to more directly show the differences in generalizability indices across designs and between groups.

Table 12 uses the D study results from the theoretically most appropriate multivariate $p^{\bullet} \times (i^{\circ} : h^{\bullet})$ design as a baseline and indicates the bias in the D study results when other designs are used. The largest bias is when both the passage and process facets are ignored, which resulted in an overestimation of about .014 for the non-ELL group and .013 for the ELL group.

Table 13 summarizes the passage effect as exhibited in the differences in some of the D study results when only the passage facet is ignored. Table 14 summarizes the cognitive process effect as exhibited in the differences in some of the D study results when only the process facet is ignored. The passage effect is larger than the process effect across the board.

Table 15 is a summary of the non-ELL minus ELL group differences in the D study results. Results for all designs consistently suggest a decrease of about .067 in both indices for the ELL group.

Conclusions and Discussion

Passage and process

Through separate analyses, this study demonstrated the differences of results when one or more of the facets of a reading comprehension test are neglected in generalizability analyses. The comparison of results across different designs suggested that the passage facet had a larger impact than the process facet for the particular reading comprehension test. Without considering the passage effect, test reliability can be overestimated. The overestimation is due to the fact that $\hat{\sigma}^2(pH)$ which contributes to error in the $p \times (I : H)$ analysis is effectively taken as part of the universe score variance in a $p \times i$ analysis.

The amount of overestimation due to neglecting the passage facet is about .015 in this study, which is consistent with findings from Wan and Brennan (2005), a study conducted for the same test using different samples. However, much larger amounts of overestimation have been reported in several other studies for the same or different reading comprehension tests. For example, Lee and Frisbie (1999) reported an overestimation of about .04 for an earlier form of the ITBS reading comprehension test. Also, an overestimation of about .10 was reported by Sireci et al (1991) for the Scholastic Aptitude Test (SAT)-Verbal test; an overestimation of about .08 was reported by Wainer (1995) for the Reading comprehension test of the Law School Admission Test; and an average overestimation of .03 was reported by Wainer and Thissen (1996) for 18 forms of the North Carolina End-of-Grade reading test. Clearly, the amount of overestimation differs from test to test. However, care should always be taken when important decisions are to be made.

In addition, passage effect also manifested in the conditional SEM. When passages were ignored, the CSEM plots were very similar between the two groups (Figures 3 and 4). When passages were taken into account, however, differences occurred (Figures 5 and 6). The smaller CSEM for the ELL group at the lower end of the score scale seems to indicates that the item difficulties within passages may have smaller variability for the ELL group than for the non-ELL group.

The amount of increase in the generalizabiility indices by considering the process effect is about .001 on average, which probably is negligible for most practical considerations. The similarities between plots of CSEM of the multivariate designs and their univariate counterparts also confirmed the negligibility of the process effect. The lack of influence of the fixed process facet on generalizability in this study may be due to the extremely high correlations between the three cognitive processes. When variables are so highly correlated, average within variable variances may not be much larger than average between variable covariances, and consequently the coefficients estimated may not be larger than when the fixed facet is ignored (Brennan, 2001).

An additional issue raised by the extremely high correlations between cognitive processes from this study is the nature the cognitive processes. Such results seem support the unitary view of reading comprehension mentioned at the beginning of the paper. However, it is always possible that the processes are conceptually distinct but the items used to measure them do so in an imperfect manner, which calls for more inquiry into the actual cognitive processes involved in answering the test questions is needed. Some recently proposed procedures might be helpful in doing this (Gierl, M. J., Tan, X., & Wang, C., 2005; Tatsuoka, K. K. 1995; Leighton, J. P., Gierl, M. J., & Hunka, S., 2004).

Between groups

The results demonstrated that test scores from the ELL group should be interpreted with less confidence than those from the non-ELL group. The univariate $p \times (i:h)$ D studies suggest that giving the ELL students a test double the length of the present one may allow us to interpret their scores with equal confidence as the non-ELL students. However, this is not likely to be practical in most occasions. Even if it is possible, a simple increase of test length may not have the anticipated effect. Indeed, sometimes a shorter test may be more reliable for this group (Abedi, personal communication). The key is to eliminate construct irrelevant distractions for the ELL students in the passages and items, which requires substantive work by content experts who are familiar with the cultural and educational background of the ELL students. The results from this study only give a sense of the magnitude of the gap between score generalizability for the two groups. How to close the gap is a much more challenging question that has no simple answer.

Through the comparison of results between groups, this study showed that generalizability theory can be very useful in demonstrating group differences. It permits a between-group detailed comparison of various sources of score variability, the relationship among variables, and conditional standard errors of measurement. For example, the univariate $p \times (i:h)$ analysis shows that passages have a larger variability for ELL students than for non-ELL students, while the multivariate $p^{\bullet} \times (i^{\circ}:h^{\bullet})$ analysis further shows that the larger variability for the ELL students mainly comes from one of the three cognitive processes.

In this paper, the across-group comparisons of variance components that contribute to error were based on the differential magnitude of the variance component. These comparisons may be misleading since the universe score variances between groups are different. Standardizing the variance components relative to the universe score variances within groups may be a better way to compare the relative contribution of each variance component between groups. Table 16 does so by reporting all variance components as ratios of universe score variance within groups. One more commonly used way of standardizing the variances is to report each variance component as a percentage of the overall variance. Yet caution should be taken that neither the ratios nor percentages be misinterpreted as indicators of the "importance" of each facet (Brennan, 2001a).

References:

Abedi, J. (2002). Psychometric issues in the assessment of English Language Learners. Presentation made at the CRESS 2002 Annual Conference: Research Goes to School, Assessment, Accountability, and Improvement. Available at http://cresst96.cse.ucla.edu/products/overheads/conf2002/overhd02/CRESST2002Abedi.ppt

Abedi, J., Courtney, M. & Leon, S. (2003). Research-supported accommodation for English Language Learners in NAEP. (CSE Technical Report No. 586). Los Angeles: UCLA Center for the Study of Evaluation (CSE)/ National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Lord, C, & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance*. (CSE Technical Report No. 478). Los Angeles: UCLA Center for the Study of Evaluation (CSE)/ National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Alderson, J. C. (1990a). Testing reading comprehension skills (Part One). Reading in a Foreign Language, 6(2), 425-438.

Alderson, J. C. (1990b). Testing reading comprehension skills (Part One). Reading in a Foreign Language, 7(1), 465-503.

Bardlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. Psychometrika, 64, 153-168.

Bloom, B. S. (1956). Taxonomy of educational objectives, Book 1: Cognitive domain. London: Longman.

Brennan, R. L. (2001a). *Generalizability Theory*. Springer.

Brennan, R. L. (2001b). Manual for mGENOVA: Version 2.1. Iowa Testing Programs occasional papers. No. 50.

Butler, F. A., & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: current trends and old dilemmas. *Language Testing*, 18 (4), pp. 409-427.

Chen, Y., Gorian, J., & Thompson, M. (2006). Verification of cognitive attributes required to solve the TIMSS-1999 mathematics items fro Taiwanese students. Paper presented at the Annual Meeting of the American Educational Research Association. April 7-11. San Francisco, CA.

Davis, F. B. (1968). Research in comprehension in reading. Reading Research Quarterly, 3, 499-545.

Gierl, M. J., Tan, X., & Wang, C. (2005). Identifying content and cognitive dimensions on the SAT®. College Board Research Report No. 2005-11. The College Board, New York.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test composed testlets. *Applied Measurement in Education*, 12(3), 237-255.

Lee, G., Dunbar, S. B., & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores from tests composed of testlets. Educational and Psychological Measurement, 61(6), 958-975.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. Journal of Educational Measurement, 41, 205-237.

Li, Yanmei, Bolt, Daniel M., & Fu, Jianbin. (2006). A comparison of alternative models for testlets. Applied Psychological Measurement, 30(1). 3-21.

Mao, X. & Li, D. (2003). Comparing English Language Learners (ELLs)'and Non-ELLs' performance on large-scale standardized content assessment. Paper presented at the fifth Annual Conference of the Midwest Association of Language Testers. West Lafayette, Indiana.

Rost, D. H. (1993). Assessing the different components of reading comprehension: Fact or fiction. Language Testing Journal, 10(1), 79-92.

Solano-Flores, G. & Trumbull, E. (2003). Examining language in context: the need for new research and practice paradigms in the testing of English-Language Learners. *Educational Researcher*, 32 (2), pp. 3-13.

Sireci, S., G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28, 237-247.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: a statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman., & R. L. Brennan (Eds.), Cognitively Diagnostic Assessment. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wan, L., & Brennan, R. L. (2005). Reliability of scores for tests composed of testlets: a comparison of approaches in three measurement models. Paper presented at the Annual Meeting of the National Council on Measurement in Education. April 12-14. Montreal, Canada.

Wan, L., Li, D. & Dunbar, S. (2006). Construct Invariance of Achievement Assessments across English Proficiency Groups: a Confirmatory Factor Analysis (CFA) Study. Paper presented at the Annual Meeting of the American Educational Research Association. April 7-11. San Francisco, CA.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. Applied Measurement in Education, 8, 157-186.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? Educational Measurement: Issues and Practices, 15(1), 22-29.

Figure 1

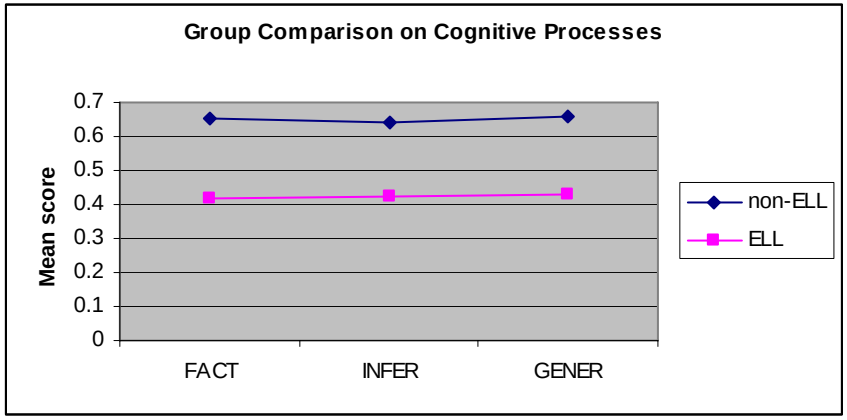

Figure 2

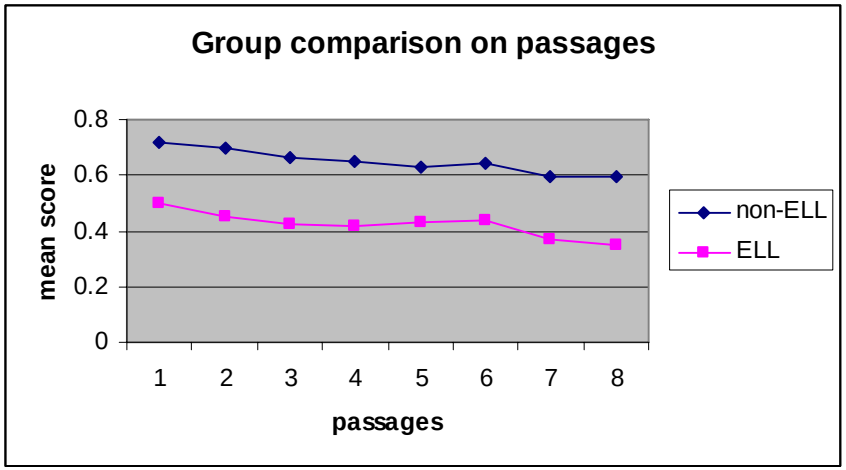

Figure 3

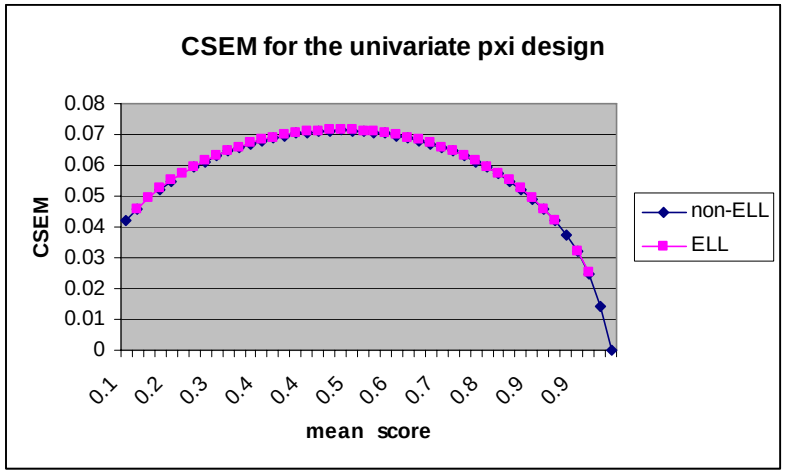

Figure 4

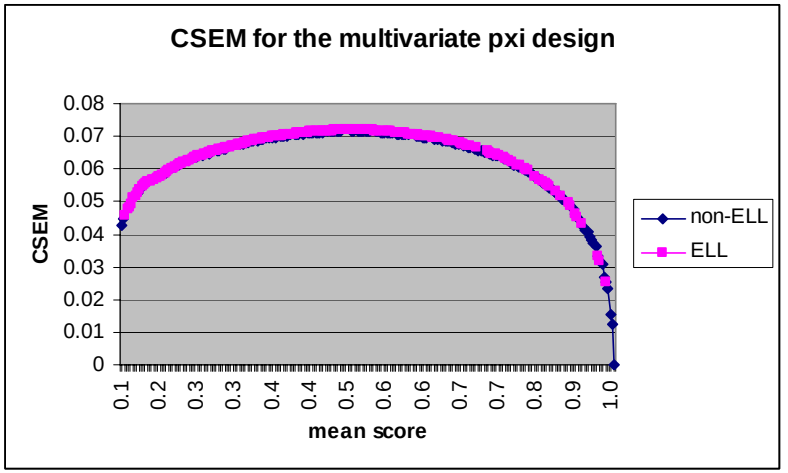

Figure 5

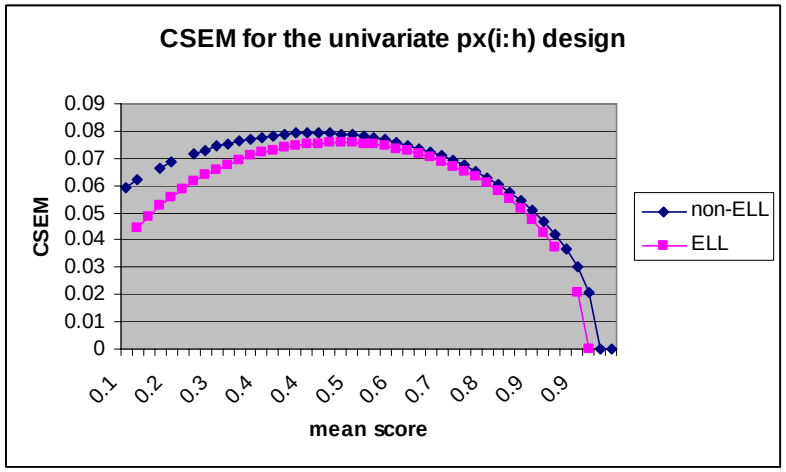

Figure 6

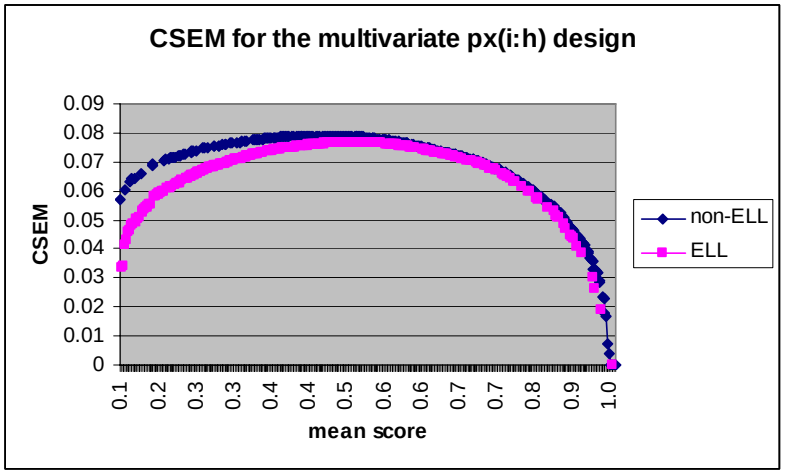

Table 1: Item distribution

	FACT	INFER	GENER	total
Passage 1	2	3	2	7
Passage 2	2	4	1	7
Passage 3	2	2	1	5
Passage 4	1	2	1	4
Passage 5	4	2	1	7
Passage 6	3	1	2	6
Passage 7	2	3	2	7
Passage 8	1	0	4	5
total	17	17	14	48

Table 2: G study and D study results of the univariate *p* x *i* design

Non-ELL			ELL		
	D study results			D study Results	
G study results	n_h'	$= n_h$	G study results	n_h'	$= n_h$
$\hat{\sigma}^2(\alpha)$	$n_{i:h}'$	$= n_{i:h}$	$\hat{\sigma}^2(\alpha)$	$n_{i:h}'$	$= n_{i:h}$
$\hat{\sigma}^2(p)$ =0.04070	$\hat{\sigma}^2(p)$	0.04070	$\hat{\sigma}^2(p)$ =0.02514	$\hat{\sigma}^2(p)$	0.02514
$\hat{\sigma}^2(i)$ = 0.01124	$\hat{\sigma}^2(I)$	0.00023	$\hat{\sigma}^2(i)$ =0.01025	$\hat{\sigma}^2(I)$	0.00021
$\hat{\sigma}^2(pi)$ =0.17596	$\hat{\sigma}^2(pI)$	0.00367	$\hat{\sigma}^2(pi)$ =0.20903	$\hat{\sigma}^2(pI)$	0.00435
	$\hat{\sigma}^2(\tau)$	0.04070		$\hat{\sigma}^2(\tau)$	0.02514
	$\hat{\sigma}^2(\delta)$	0.00367		$\hat{\sigma}^2(\delta)$	0.00435
	$\hat{\sigma}^2(\Delta)$	0.00390		$\hat{\sigma}^2(\Delta)$	0.00457
	$E\hat{\rho}^2$	0.91737		$E\hat{\rho}^2$	0.85237
	$\hat{\Phi}$	0.91256		$\hat{\Phi}$	0.84625

Table 3 : G study results of the p x i multivariate design

	Mean squares (diagonal**), observed covariances (**lower diagonal**) and observed correlations (**upper diagonal**)**			**Estimated variance (**diagonal**), covariances (**lower diagonal**) and disattenuated correlations (**upper diagonal**)**		
	Non-ELL		ELL	Non-ELL		ELL
p	$\begin{bmatrix} 0.82775 & 0.79625 & 0.80801 \\ 0.03928 & 0.84949 & 0.74262 \\ 0.04305 & 0.04009 & 0.81632 \end{bmatrix}$		$\begin{bmatrix} 0.66044 & 0.68776 & 0.66094 \\ 0.02519 & 0.58681 & 0.64191 \\ 0.02646 & 0.02422 & 0.57737 \end{bmatrix}$	$\begin{bmatrix} 0.03856 & 1.00903 & 1.02265 \\ 0.03928 & 0.03930 & 0.94317 \\ 0.04305 & 0.04009 & 0.04597 \end{bmatrix}$		$\begin{bmatrix} 0.02682 & 1.03723 & 0.99775 \\ 0.02519 & 0.02199 & 1.00875 \\ 0.02646 & 0.02422 & 0.02622 \end{bmatrix}$
i	$\begin{bmatrix} 8.66184 & & \\ & 5.18709 & \\ & & 3.78815 \end{bmatrix}$		$\begin{bmatrix} 6.68694 & & \\ & 4.94681 & \\ & & 4.92730 \end{bmatrix}$	$\begin{bmatrix} 0.01701 & & \\ & 0.01003 & \\ & & 0.00725 \end{bmatrix}$		$\begin{bmatrix} 0.01296 & & \\ & 0.00947 & \\ & & 0.00943 \end{bmatrix}$
pi	$\begin{bmatrix} 0.17228 & & \\ & 0.18147 & \\ & & 0.17276 \end{bmatrix}$		$\begin{bmatrix} 0.20456 & & \\ & 0.21303 & \\ & & 0.21032 \end{bmatrix}$	$\begin{bmatrix} 0.17228 & & \\ & 0.18147 & \\ & & 0.17276 \end{bmatrix}$		$\begin{bmatrix} 0.20456 & & \\ & 0.21303 & \\ & & 0.21032 \end{bmatrix}$

Table 4 : Default D study results of the p x i multivariate design (sample size weights)

	Non-ELL				ELL			
	FACT	INFER	GENER	Composite	FACT	INFER	GENER	Composite
$\hat{\sigma}^2(\tau)$	0.03856	0.03930	0.04597	0.04071	0.02682	0.02199	0.02622	0.02514
$\hat{\sigma}^2(\delta)$	0.01013	0.01067	0.01234	0.00366	0.01203	0.01253	0.01502	0.00436
$\hat{\sigma}^2(\Delta)$	0.01113	0.01126	0.01286	0.00390	0.01280	0.01309	0.01570	0.00458
$E\hat{\rho}^2$	0.79187	0.78638	0.78836	0.91751	0.69026	0.63696	0.63572	0.85222
$\hat{\Phi}$	0.77593	0.77720	0.78143	0.91250	0.67697	0.62685	0.62550	0.84583
Contributions to								
$\hat{\sigma}^2(\tau)$	34.91%	34.38%	30.71%		36.82%	33.49%	29.69%	
$\hat{\sigma}^2(\delta)$	34.73%	36.58%	28.68%		34.62%	36.06%	29.32%	
$\hat{\sigma}^2(\Delta)$	35.78%	36.20%	28.02%		35.03%	35.83%	29.14%	

Table 5: Some D study results (change in sample size and weights) of the multivariate pxi design for the non-ELL group

	Same sample sizes ($n_1 = 17, n_2 = 17, n_3 = 14$) with equal w-weights (.3333)				**Different sample sizes ($n_1 = n_2 = n_3 = 25$) with equal w-weights (.3333)**			
	FACT	INFER	GENER	Composite	FACT	INFER	GENER	Composite
$\hat{\sigma}^2(\tau)$	0.03856	0.03930	0.04597	0.04096	0.03856	0.03930	0.04597	0.04096
$\hat{\sigma}^2(\delta)$	0.01013	0.01067	0.01234	0.00368	0.00689	0.00726	0.00691	0.00234
$\hat{\sigma}^2(\Delta)$	0.01113	0.01126	0.01286	0.00392	0.00757	0.00766	0.00720	0.00249
$E\hat{\rho}^2$	0.79187	0.78638	0.78836	0.91750	0.84838	0.84408	0.86931	0.94596
$\hat{\Phi}$	0.77593	0.77720	0.78143	0.91271	0.83586	0.83687	0.86458	0.94264
Contributions to								
$\hat{\sigma}^2(\tau)$	32.79%	32.19%	35.02%		32.79%	32.19%	35.02%	
$\hat{\sigma}^2(\delta)$	30.57%	32.20%	37.23%		32.72%	34.47%	32.81%	
$\hat{\sigma}^2(\Delta)$	31.58%	31.95%	36.47%		33.75%	34.15%	32.10%	

Table 6: Some D study results (change in sample size and weights) of the multivariate pxi design for the ELL group

	Same sample sizes ($n_1 = 17, n_2 = 17, n_3 = 14$) with equal w-weights (.3333)				**Different sample sizes ($n_1 = n_2 = n_3 = 25$) with equal w-weights (.3333)**			
	FACT	INFER	GENER	Composite	FACT	INFER	GENER	Composite
$\hat{\sigma}^2(\tau)$	0.02682	0.02199	0.02622	0.02519	0.02682	0.02199	0.02622	0.02519
$\hat{\sigma}^2(\delta)$	0.01203	0.01253	0.01502	0.00440	0.00818	0.00852	0.00841	0.00279
$\hat{\sigma}^2(\Delta)$	0.01280	0.01309	0.01570	0.00462	0.00870	0.00890	0.00879	0.00293
$E\hat{\rho}^2$	0.69026	0.63696	0.63572	0.85136	0.76621	0.72069	0.75706	0.90027
$\hat{\Phi}$	0.67697	0.62685	0.62550	0.84503	0.75502	0.71185	0.74890	0.89574
Contributions to								
$\hat{\sigma}^2(\tau)$	34.60%	31.49%	33.91%		34.60%	31.49%	33.91%	
$\hat{\sigma}^2(\delta)$	30.40%	31.65%	37.95%		32.58%	33.93%	33.50%	
$\hat{\sigma}^2(\Delta)$	30.77%	31.48%	37.75%		32.97%	33.72%	33.31%	

Table 7: G study and some D study results of the px(i:h) design for the non-ELL group

G study results	D studies	default	Maintaining 48 items			
	n_h'	$= n_h$	4	6	8	12
$\hat{\sigma}^2(\alpha)$	$n_{i:h}'$	$= n_{i:h}$	12	8	6	4
$\hat{\sigma}^2(p)$=0.04009	$\hat{\sigma}^2(p)$	0.04009	0.04009	0.04009	0.04009	0.04009
$\hat{\sigma}^2(h)$ = 0.00007	$\hat{\sigma}^2(H)$	0.00001	0.00002	0.00001	0.00001	0.00000
$\hat{\sigma}^2(i:h)$=0.01108	$\hat{\sigma}^2(I:H)$	0.00023	0.00023	0.00023	0.00023	0.00023
$\hat{\sigma}^2(ph)$=0.00642	$\hat{\sigma}^2(pH)$	0.00083	0.00161	0.00107	0.00080	0.00054
$\hat{\sigma}^2(pi:h)$=0.17006	$\hat{\sigma}^2(pI:H)$	0.00354	0.00354	0.00354	0.00354	0.00354
	$\hat{\sigma}^2(\tau)$	0.04009	0.04009	0.04009	0.04009	0.04009
	$\hat{\sigma}^2(\delta)$	0.00437	0.00515	0.00461	0.00435	0.00408
	$\hat{\sigma}^2(\Delta)$	0.00461	0.00540	0.00486	0.00458	0.00431
	$E\hat{\rho}^2$	0.90165	0.88621	0.89682	0.90221	0.90768
	$\hat{\Phi}$	0.89681	0.88138	0.89198	0.89738	0.90284

Table 8: G study and some D study results of the p x (i:h) design for the ELL group:

	D studies	Default	Maintaining 48 items				Some other passage and item combinations			
G study results	n_h'	$= n_h$	4	6	8	12	8	9	10	15
$\hat{\sigma}^2(\alpha)$	$n_{i:h}'$	$= n_{i:h}$	12	8	6	4	7	7	6	6
$\hat{\sigma}^2(p)$=0.02471	$\hat{\sigma}^2(p)$	0.02471	0.02471	0.02471	0.02471	0.02471	0.02471	0.02471	0.02471	0.02471
$\hat{\sigma}^2(h)$=0.00048	$\hat{\sigma}^2(H)$	0.00006	0.00012	0.00008	0.00006	0.00004	0.00006	0.00005	0.00005	0.00003
$\hat{\sigma}^2(i:h)$=0.00982	$\hat{\sigma}^2(I:H)$	0.00020	0.00020	0.00020	0.00020	0.00020	0.00018	0.00005	0.00016	0.00011
$\hat{\sigma}^2(ph)$=0.00388	$\hat{\sigma}^2(pH)$	0.00050	0.00097	0.00065	0.00049	0.00032	0.00049	0.00043	0.00039	0.00026
$\hat{\sigma}^2(pi:h)$=0.20558	$\hat{\sigma}^2(pI:H)$	0.00428	0.00428	0.00428	0.00428	0.00428	0.00367	0.00326	0.00343	0.00228
	$\hat{\sigma}^2(\tau)$	0.02471	0.02471	0.02471	0.02471	0.02471	0.02471	0.02471	0.02471	0.02471
	$\hat{\sigma}^2(\delta)$	0.00479	0.00525	0.00493	0.00477	0.00461	0.00416	0.00369	0.00381	0.00254
	$\hat{\sigma}^2(\Delta)$	0.00505	0.00558	0.00522	0.00503	0.00485	0.06627	0.00390	0.00403	0.00268
	$E\hat{\rho}^2$	0.83778	0.82468	0.83368	0.83826	0.84289	0.85602	0.86994	0.86628	0.90670
	$\hat{\Phi}$	0.83026	0.81582	0.82574	0.83079	0.83591	0.84909	0.86357	0.85989	0.90202

Table 9 : G study results of the multivariate $p^{\bullet} \times (i^{\circ} : h^{\bullet})$ design

	Estimated variances (diagonal**), covariances (**lower diagonal**) and disattenuated correlations (**upper diagonal**)**	
	Non-ELL	ELL
p	$\begin{bmatrix} 0.03823 & 1.02890 & 1.04553 \\ 0.03859 & 0.03680 & 1.03066 \\ 0.04358 & 0.04215 & 0.04544 \end{bmatrix}$	$\begin{bmatrix} 0.02714 & 1.05070 & 1.09834 \\ 0.02530 & 0.02137 & 1.08757 \\ 0.02779 & 0.02442 & 0.02358 \end{bmatrix}$
h	$\begin{bmatrix} 0.00079 & & \\ 0.00128 & 0.00000 & \\ -0.00149 & -0.00001 & 0.00023 \end{bmatrix}$	$\begin{bmatrix} 0.00000 & & \\ 0.00105 & 0.00102 & \\ -0.00071 & -0.00039 & 0.00357 \end{bmatrix}$
i:h	$\begin{bmatrix} 0.01623 & & \\ & 0.01068 & \\ & & 0.00868 \end{bmatrix}$	$\begin{bmatrix} 0.01400 & & \\ & 0.00765 & \\ & & 0.00866 \end{bmatrix}$
ph	$\begin{bmatrix} 0.00211 & & \\ 0.00938 & 0.00750 & \\ 0.00520 & 0.00444 & 0.00457 \end{bmatrix}$	$\begin{bmatrix} 0.00360 & & \\ 0.00312 & 0.00365 & \\ 0.00344 & 0.00432 & 0.00989 \end{bmatrix}$
pi:h	$\begin{bmatrix} 0.17261 & & \\ & 0.17459 & \\ & & 0.17123 \end{bmatrix}$	$\begin{bmatrix} 0.20228 & & \\ & 0.21229 & \\ & & 0.20084 \end{bmatrix}$

Table 10: Default D study results of the multivariate $p^{\bullet} \times (i^{\circ} : h^{\bullet})$ design (sample size weights)

	Non-ELL				**ELL**			
	FACT	INFER	GENER	Composite	FACT	INFER	GENER	Composite
$\hat{\sigma}^2(\tau)$	0.03823	0.03680	0.04544	0.04067	0.02714	0.02137	0.02358	0.02522
$\hat{\sigma}^2(\delta)$	0.01047	0.01149	0.01298	0.00437	0.01243	0.01308	0.01596	0.00483
$\hat{\sigma}^2(\Delta)$	0.01154	0.01201	0.01363	0.00464	0.01326	0.01370	0.01716	0.00512
$E\hat{\rho}^2$	0.78506	0.76209	0.77785	0.90290	0.68577	0.62030	0.59638	0.83931
$\hat{\Phi}$	0.76806	0.75400	0.76918	0.89768	0.67179	0.60942	0.57880	0.83126
Contributions to								
$\hat{\sigma}^2(\tau)$	34.76%	33.96%	31.28%		37.46%	33.21%	29.33%	
$\hat{\sigma}^2(\delta)$	34.86%	37.45%	27.69%		34.17%	36.95%	29.88%	
$\hat{\sigma}^2(\Delta)$	35.87%	37.18%	26.95%		34.41%	35.66%	29.94%	

Table 11: Comparison of default D study results of different designs across groups

	Non-ELL				ELL			
	$p \times i$	$p^{\bullet} \times i^{\circ}$	$p \times (i:h)$	$p^{\bullet} \times (i^{\circ}:h^{\bullet})$	$p \times i$	$p^{\bullet} \times i^{\circ}$	$p \times (i:h)$	$p^{\bullet} \times (i^{\circ}:h^{\bullet})$
$\hat{\sigma}^2(\tau)$	0.04070	0.04071	0.04009	0.04067	0.02514	0.02514	0.02471	0.02522
$\hat{\sigma}^2(\delta)$	0.00367	0.00366	0.00437	0.00437	0.00435	0.00436	0.00479	0.00483
$\hat{\sigma}^2(\Delta)$	0.00390	0.00390	0.00461	0.00464	0.00457	0.00458	0.00505	0.00512
$E\hat{\rho}^2$	0.91737	0.91751	0.90165	0.90290	0.85237	0.85222	0.83778	0.83931
$\hat{\Phi}$	0.91256	0.91250	0.89681	0.89768	0.84625	0.84583	0.83026	0.83126

Table 12: Bias of the D study results from other designs compared with the $p^{\bullet} \times (i^{\circ}:h^{\bullet})$ design:

	Non-ELL			ELL		
	$p \times i$	$p^{\bullet} \times i^{\circ}$	$p \times (i:h)$	$p \times i$	$p^{\bullet} \times i^{\circ}$	$p \times (i:h)$
$\hat{\sigma}^2(\tau)$	0.00003	0.00004	-0.00058	-0.00008	-0.00008	-0.00051
$\hat{\sigma}^2(\delta)$	-0.00070	-0.00071	0.00000	-0.00048	-0.00047	-0.00046
$\hat{\sigma}^2(\Delta)$	-0.00074	-0.00074	-0.00003	-0.00055	-0.00054	-0.00007
$E\hat{\rho}^2$	0.01447	0.01461	-0.00125	0.01306	0.01291	-0.00153
$\hat{\Phi}$	0.01488	0.01482	-0.00087	0.01499	0.01457	-0.00100

Table 13: Passage effect – Differences of D study results when passages are ignored

	Non-ELL		ELL	
	$p^{\bullet} \times i^{\circ} - p^{\bullet} \times (i^{\circ} : h^{\bullet})$	$p \times i - p \times (i : h)$	$p^{\bullet} \times i^{\circ} - p^{\bullet} \times (i^{\circ} : h^{\bullet})$	$p \times i - p \times (i : h)$
$\hat{\sigma}^2(\delta)$	-0.00071	-0.00070	-0.00047	-0.00044
$\hat{\sigma}^2(\Delta)$	-0.00074	-0.00071	-0.00054	-0.00048
$E\hat{\rho}^2$	0.01461	0.01572	0.01291	0.01459
$\hat{\Phi}$	0.01482	0.01575	0.01457	0.01599

Table 14: Process effect – Differences of D study results when processes are ignored

	Non-ELL		ELL	
	$p \times i - p^{\bullet} \times i^{\circ}$	$p \times (i : h) - p^{\bullet} \times (i^{\circ} : h^{\bullet})$	$p \times i - p^{\bullet} \times i^{\circ}$	$p \times (i : h) - p^{\bullet} \times (i^{\circ} : h^{\bullet})$
$\hat{\sigma}^2(\delta)$	0.00001	0.00000	-0.00001	-0.00004
$\hat{\sigma}^2(\Delta)$	0.00000	-0.00003	-0.00001	-0.00007
$E\hat{\rho}^2$	-0.00014	-0.00125	0.00015	-0.00153
$\hat{\Phi}$	0.00006	-0.00100	0.00042	-0.00100

Table 15: non-ELL minus ELL differences in D study results in different designs

	$p \times i$	$p^{\bullet} \times i^{\circ}$	$p \times (i:h)$	$p^{\bullet} \times (i^{\circ}:h^{\bullet})$
$\hat{\sigma}^2(\tau)$	0.01556	0.01557	0.01538	0.01545
$\hat{\sigma}^2(\delta)$	-0.00068	-0.00070	-0.00042	-0.00046
$\hat{\sigma}^2(\Delta)$	-0.00067	-0.00068	-0.00044	-0.00048
$E\hat{\rho}^2$	0.06500	0.06529	0.06387	0.06359
$\hat{\Phi}$	0.06631	0.06667	0.06655	0.06642

Table 16: Ratios of variance components to the universe score components based on the results of the multivariate $p^{\bullet} \times (i^{\circ}:h^{\bullet})$ design

	Non-ELL			ELL		
	FACT	INFER	GENER	FACT	INFER	GENER
p	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
h	0.02066	0.00000	0.00506	0.00000	0.04773	0.15140
i:h	0.42454	0.29022	0.19102	0.51584	0.35798	0.36726
ph	0.05519	0.20380	0.10057	0.13265	0.17080	0.41942
pi:h	4.51504	4.47429	3.76827	7.45321	9.93402	8.51739