The George Washington University
WASHINGTON DC

Center for Equity and Excellence in Education

Test Validity and Mean Effects of Test Accommodations for ELLs and Non-ELLs: A Meta-Analysis

Maria Pennock-Roman, PhD
MPR Psychometric Research & Consulting
mpennock@maine.rr.com

Charlene Rivera, EdD
Center for Equity and Excellence in Education
The George Washington University
crivera@ceee.gwu.edu

The National Center for the Improvement of Educational Assessment, WestEd, and The Assessment and Accountability Comprehensive Center

The 2007 Reidy Interactive Lecture Series
English Language Learner Assessment and Accountability:
Critical Considerations for Design and Implementation

September 27-28, 2007

Test Validity and Mean Effects of Test Accommodations for ELLs and Non-ELLs:
A Meta-Analysis

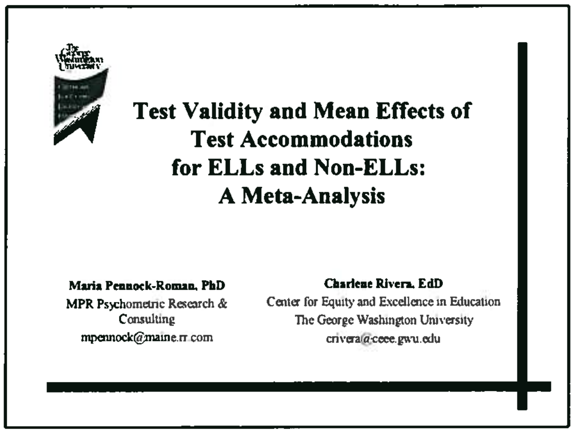
Charlene Rivera and Maria Pennock-Roman

The objective of the meta-analysis was to examine to what degree different types of accommodations lead to possible improvements in the test performance of English language learners (ELLs) on content tests such as mathematics and science. This analysis is a first step to examine whether ELLs have increased access to the content of the test with accommodations intended to reduce construct irrelevant variance due to English proficiency. We reviewed and summarized all the available U.S. studies that randomly assigned school-aged ELLs to test accommodation conditions that were paired with a control condition (the original, unaccommodated version of the test booklet). In the meta-analysis, 13 experimental studies were included. The unit of analysis was a subsample assigned to a particular accommodation paired to a corresponding control group. There were altogether 44 ELL subsamples and 32 non-ELL subsamples receiving an accommodation. We calculated Glass's *d* for each subsample of ELLs and of non-ELLs (*d* is an effect size--the difference between the accommodation and control means divided by the control group standard deviation). To find average effect sizes, subsamples were classified according to type of test accommodation, time allowed during test administration, and level of English language proficiency, where available. Although standard errors were calculated, we stress effect size magnitude in our interpretation rather than statistical significance owing to low statistical power for most accommodation categories.

For ELLs not categorized by level of English proficiency, the Pop-up Glossary in English (computer administered) was the most effective English language accommodation when the original, restricted time limits were used. Other English language accommodations and Spanish-English, bilingual accommodations, were less effective or sometimes harder than the original test booklet under restricted time conditions. There was evidence that providing generous time limits produced a differentially larger improvement for ELLs who received an accommodation as compared with ELLs who were administered the original booklet using extended time, or with non-ELLs. For example, for the Spanish Option/Dual Language booklets, the average effect size under restricted time limits was 0.003 as compared with 0.299 when control and experimental groups were both allowed to have generous time limits. Also, when extra time was provided for both the accommodated and original booklets, the two versions were generally equal in overall difficulty for non-ELLs. As a result, we recommend the following accommodations, provided that extended time is allowed for both original and accommodated booklets: English Dictionary/Glossary, Spanish Option/Dual language, and Spanish-English Dictionary/Glossary. Extended time without a linguistic accommodation was also somewhat effective, on average, but it may alter the difficulty of the test for non-ELLs and therefore change the score scale. The Small Group administration condition was not effective for ELLs.

When information about English language proficiency level was available, the Spanish test version was by far the most effective accommodation, provided that students had low proficiency in English and/or had received instruction in Spanish for the content of the test (mean effect size of +1.323). The plain or simplified English accommodation was more effective for students at intermediate levels of English proficiency, or for students receiving content instruction in English.

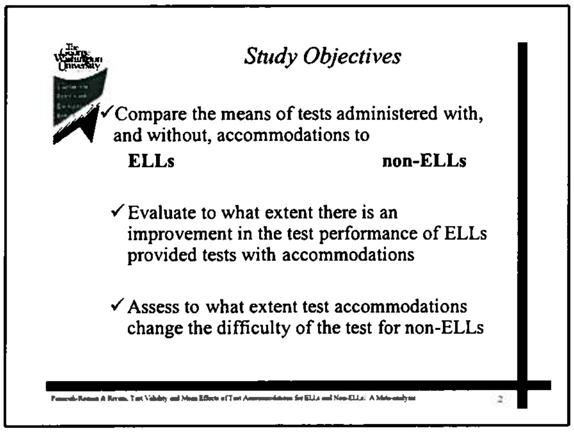
It is reasonable to assume that test accommodations are congeneric to their original test booklets owing to having the same test specifications. Nevertheless, follow-up analyses of psychometric validity of test accommodations with approaches other than means are desirable (National Research Council, 2004).



Test Validity and Mean Effects of Test Accommodations for ELLs and Non-ELLs: A Meta-Analysis

Maria Pennock-Roman, PhD
MPR Psychometric Research &
Consulting
mpennock@maine.rr.com

Charlene Rivera, EdD
Center for Equity and Excellence in Education
The George Washington University
crivera@ceee.gwu.edu



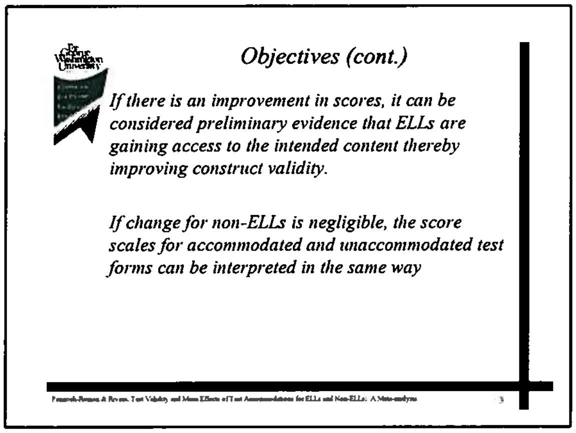
Study Objectives

✓Compare the means of tests administered with, and without, accommodations to

ELLs **non-ELLs**

✓Evaluate to what extent there is an improvement in the test performance of ELLs provided tests with accommodations

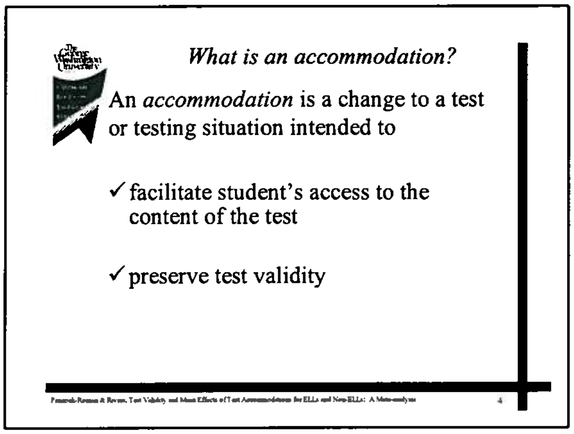
✓Assess to what extent test accommodations change the difficulty of the test for non-ELLs



Objectives (cont.)

If there is an improvement in scores, it can be considered preliminary evidence that ELLs are gaining access to the intended content thereby improving construct validity.

If change for non-ELLs is negligible, the score scales for accommodated and unaccommodated test forms can be interpreted in the same way

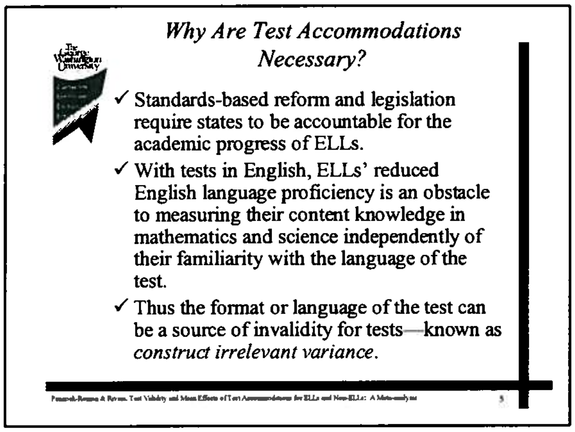


What is an accommodation?

An *accommodation* is a change to a test or testing situation intended to

✓ facilitate student's access to the content of the test

✓ preserve test validity

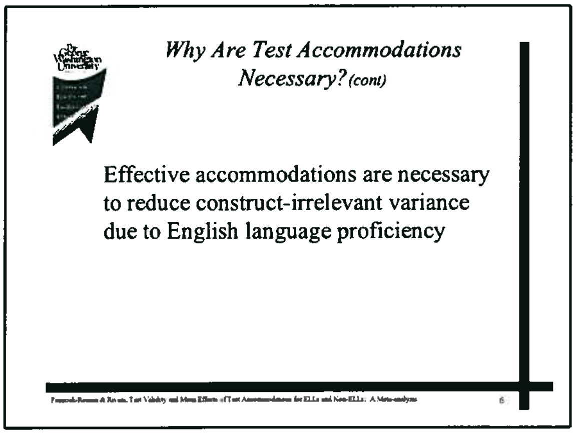


Why Are Test Accommodations Necessary?

✓ Standards-based reform and legislation require states to be accountable for the academic progress of ELLs.

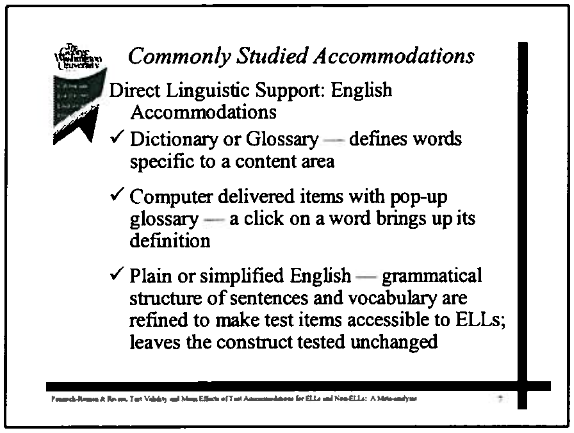
✓ With tests in English, ELLs' reduced English language proficiency is an obstacle to measuring their content knowledge in mathematics and science independently of their familiarity with the language of the test.

✓ Thus the format or language of the test can be a source of invalidity for tests—known as *construct irrelevant variance*.



Why Are Test Accommodations Necessary? (cont)

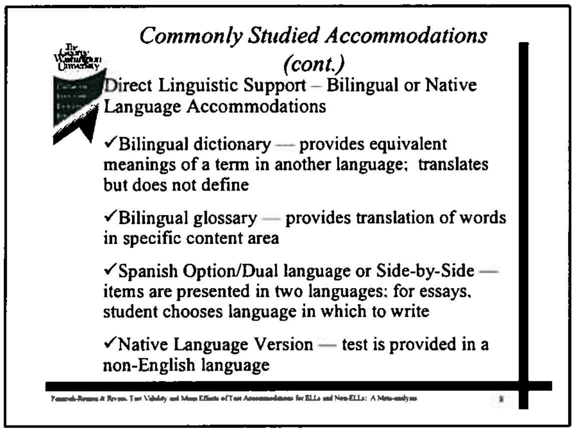
Effective accommodations are necessary to reduce construct-irrelevant variance due to English language proficiency



Commonly Studied Accommodations

Direct Linguistic Support: English Accommodations

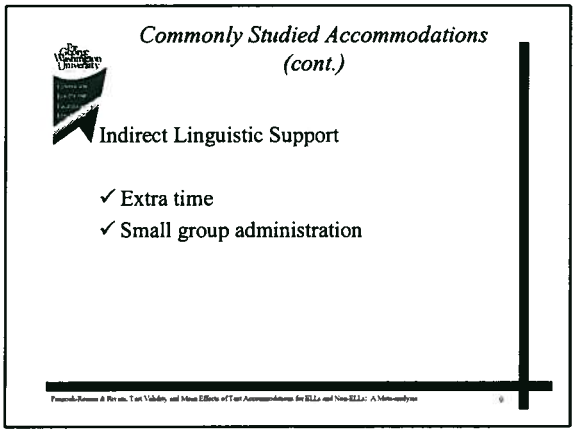
- ✓ Dictionary or Glossary — defines words specific to a content area
- ✓ Computer delivered items with pop-up glossary — a click on a word brings up its definition
- ✓ Plain or simplified English — grammatical structure of sentences and vocabulary are refined to make test items accessible to ELLs; leaves the construct tested unchanged



Commonly Studied Accommodations (cont.)

Direct Linguistic Support – Bilingual or Native Language Accommodations

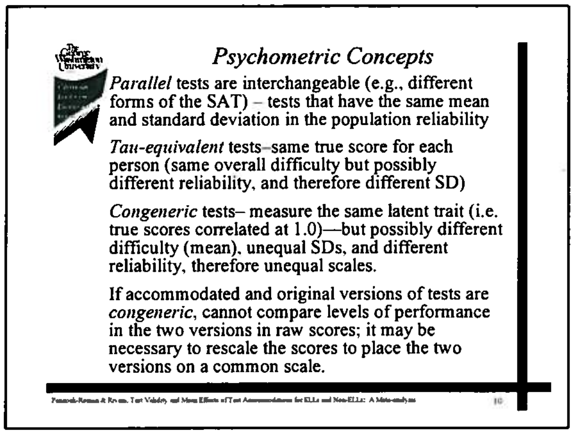
- ✓Bilingual dictionary — provides equivalent meanings of a term in another language; translates but does not define
- ✓Bilingual glossary — provides translation of words in specific content area
- ✓Spanish Option/Dual language or Side-by-Side — items are presented in two languages: for essays, student chooses language in which to write
- ✓Native Language Version — test is provided in a non-English language



Commonly Studied Accommodations (cont.)

Indirect Linguistic Support

- ✓ Extra time
- ✓ Small group administration



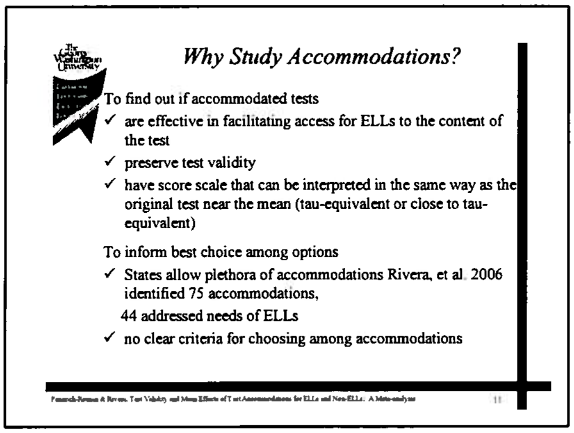
Psychometric Concepts

Parallel tests are interchangeable (e.g., different forms of the SAT) – tests that have the same mean and standard deviation in the population reliability

Tau-equivalent tests–same true score for each person (same overall difficulty but possibly different reliability, and therefore different SD)

Congeneric tests– measure the same latent trait (i.e. true scores correlated at 1.0)—but possibly different difficulty (mean), unequal SDs, and different reliability, therefore unequal scales.

If accommodated and original versions of tests are *congeneric*, cannot compare levels of performance in the two versions in raw scores; it may be necessary to rescale the scores to place the two versions on a common scale.



Why Study Accommodations?

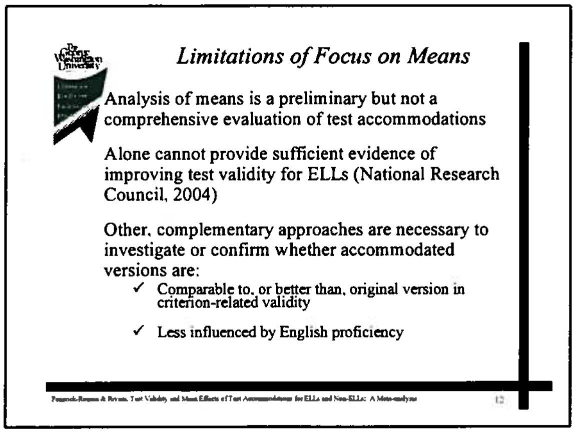
To find out if accommodated tests

- ✓ are effective in facilitating access for ELLs to the content of the test
- ✓ preserve test validity
- ✓ have score scale that can be interpreted in the same way as the original test near the mean (tau-equivalent or close to tau-equivalent)

To inform best choice among options

- ✓ States allow plethora of accommodations Rivera, et al. 2006 identified 75 accommodations,

  44 addressed needs of ELLs
- ✓ no clear criteria for choosing among accommodations



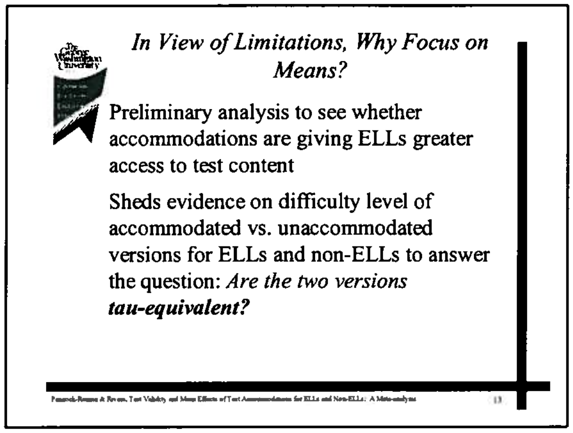
Limitations of Focus on Means

Analysis of means is a preliminary but not a comprehensive evaluation of test accommodations

Alone cannot provide sufficient evidence of improving test validity for ELLs (National Research Council, 2004)

Other, complementary approaches are necessary to investigate or confirm whether accommodated versions are:

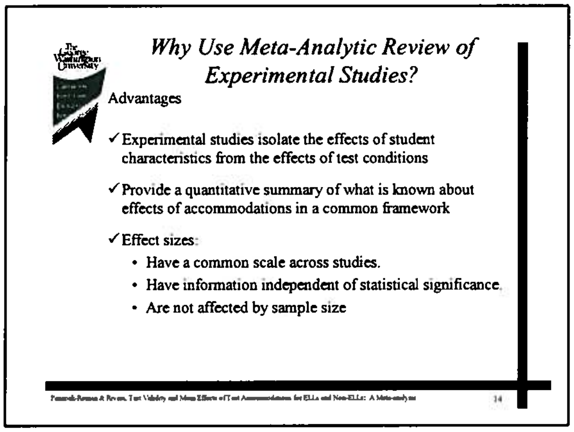
- ✓ Comparable to, or better than, original version in criterion-related validity
- ✓ Less influenced by English proficiency



In View of Limitations, Why Focus on Means?

Preliminary analysis to see whether accommodations are giving ELLs greater access to test content

Sheds evidence on difficulty level of accommodated vs. unaccommodated versions for ELLs and non-ELLs to answer the question: *Are the two versions* ***tau-equivalent?***



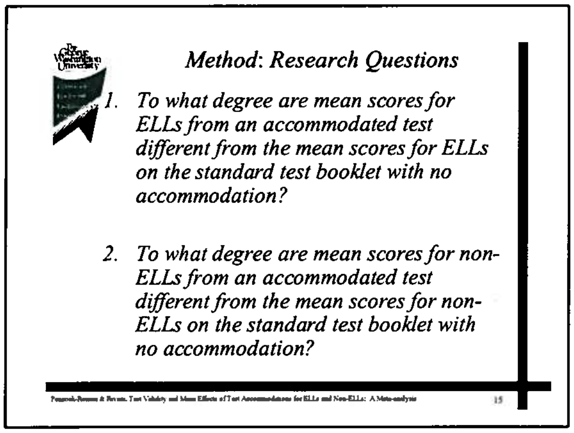
Why Use Meta-Analytic Review of Experimental Studies?

Advantages

✓ Experimental studies isolate the effects of student characteristics from the effects of test conditions

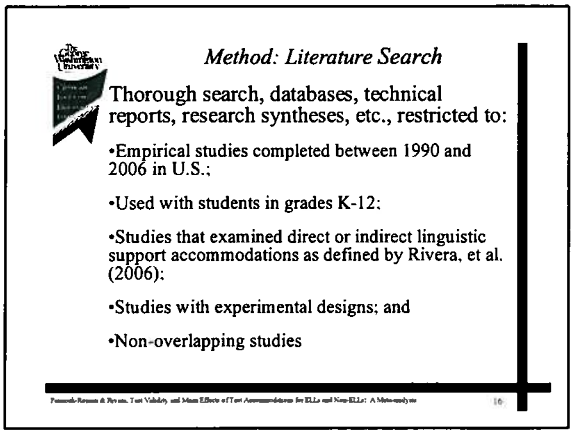
✓ Provide a quantitative summary of what is known about effects of accommodations in a common framework

✓ Effect sizes:
- Have a common scale across studies.
- Have information independent of statistical significance
- Are not affected by sample size



Method: Research Questions

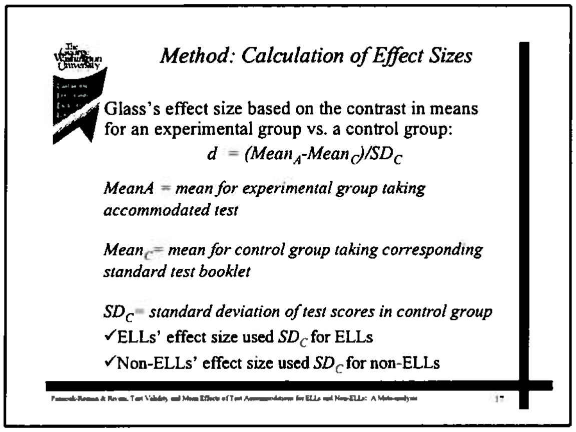
1. *To what degree are mean scores for ELLs from an accommodated test different from the mean scores for ELLs on the standard test booklet with no accommodation?*
2. *To what degree are mean scores for non-ELLs from an accommodated test different from the mean scores for non-ELLs on the standard test booklet with no accommodation?*



Method: Literature Search

Thorough search, databases, technical reports, research syntheses, etc., restricted to:

- Empirical studies completed between 1990 and 2006 in U.S.;
- Used with students in grades K-12;
- Studies that examined direct or indirect linguistic support accommodations as defined by Rivera, et al. (2006);
- Studies with experimental designs; and
- Non-overlapping studies



Method: Calculation of Effect Sizes

Glass's effect size based on the contrast in means for an experimental group vs. a control group:

$$d = (Mean_A - Mean_C)/SD_C$$

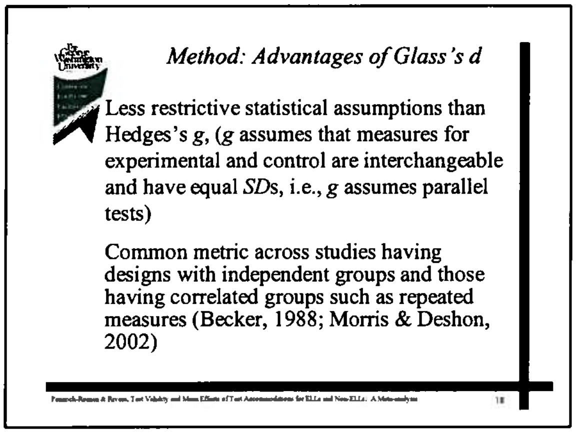
MeanA = mean for experimental group taking accommodated test

$Mean_C$= mean for control group taking corresponding standard test booklet

SD_C= standard deviation of test scores in control group

✓ELLs' effect size used SD_C for ELLs

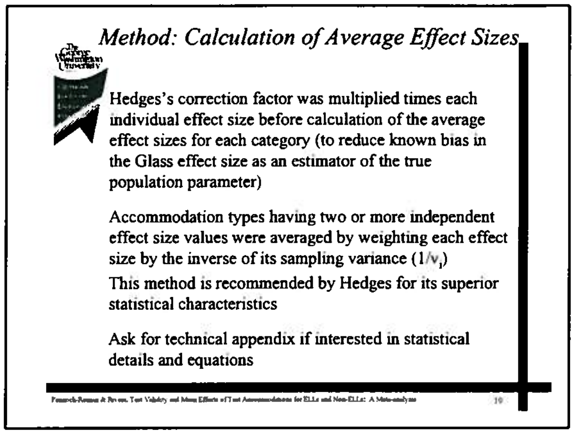
✓Non-ELLs' effect size used SD_C for non-ELLs



Method: Advantages of Glass's d

Less restrictive statistical assumptions than Hedges's *g*, (*g* assumes that measures for experimental and control are interchangeable and have equal *SD*s, i.e., *g* assumes parallel tests)

Common metric across studies having designs with independent groups and those having correlated groups such as repeated measures (Becker, 1988; Morris & Deshon, 2002)



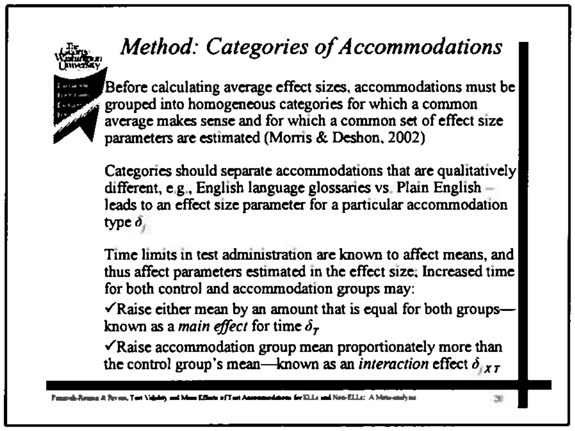
Method: Calculation of Average Effect Sizes

Hedges's correction factor was multiplied times each individual effect size before calculation of the average effect sizes for each category (to reduce known bias in the Glass effect size as an estimator of the true population parameter)

Accommodation types having two or more independent effect size values were averaged by weighting each effect size by the inverse of its sampling variance ($1/v_i$)

This method is recommended by Hedges for its superior statistical characteristics

Ask for technical appendix if interested in statistical details and equations



Method: Categories of Accommodations

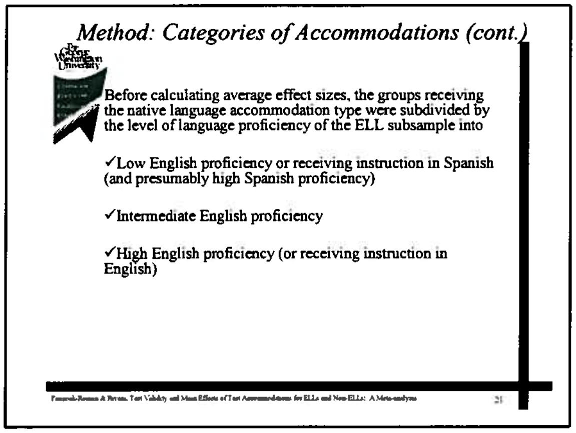
Before calculating average effect sizes, accommodations must be grouped into homogeneous categories for which a common average makes sense and for which a common set of effect size parameters are estimated (Morris & Deshon, 2002)

Categories should separate accommodations that are qualitatively different, e.g., English language glossaries vs. Plain English – leads to an effect size parameter for a particular accommodation type δ_j

Time limits in test administration are known to affect means, and thus affect parameters estimated in the effect size; Increased time for both control and accommodation groups may:

✓Raise either mean by an amount that is equal for both groups—known as a *main effect* for time δ_T

✓Raise accommodation group mean proportionately more than the control group's mean—known as an *interaction* effect $\delta_{j\,X\,T}$



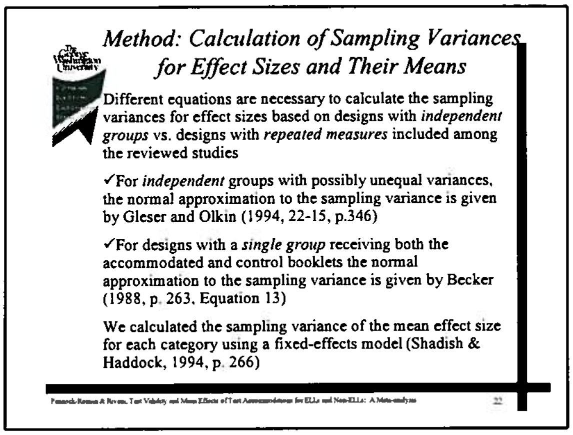
Method: Categories of Accommodations (cont.)

Before calculating average effect sizes, the groups receiving the native language accommodation type were subdivided by the level of language proficiency of the ELL subsample into

✓Low English proficiency or receiving instruction in Spanish (and presumably high Spanish proficiency)

✓Intermediate English proficiency

✓High English proficiency (or receiving instruction in English)



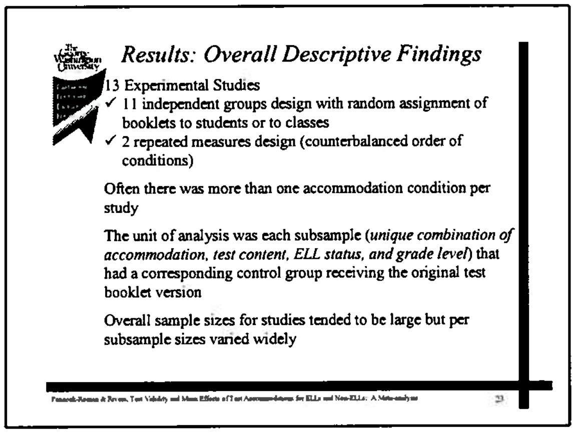
Method: Calculation of Sampling Variances for Effect Sizes and Their Means

Different equations are necessary to calculate the sampling variances for effect sizes based on designs with *independent groups* vs. designs with *repeated measures* included among the reviewed studies

✓For *independent* groups with possibly unequal variances, the normal approximation to the sampling variance is given by Gleser and Olkin (1994, 22-15, p.346)

✓For designs with a *single group* receiving both the accommodated and control booklets the normal approximation to the sampling variance is given by Becker (1988, p. 263, Equation 13)

We calculated the sampling variance of the mean effect size for each category using a fixed-effects model (Shadish & Haddock, 1994, p. 266)



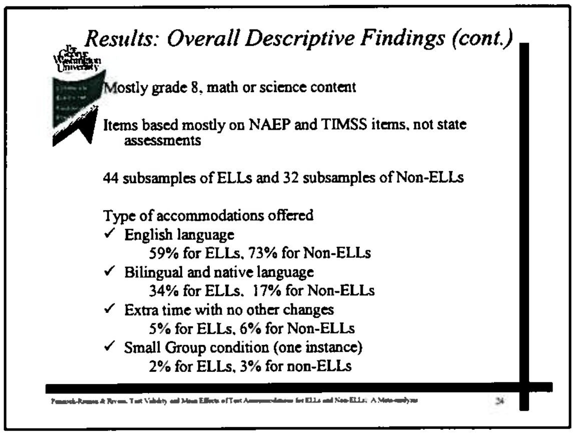
Results: Overall Descriptive Findings

13 Experimental Studies
- ✓ 11 independent groups design with random assignment of booklets to students or to classes
- ✓ 2 repeated measures design (counterbalanced order of conditions)

Often there was more than one accommodation condition per study

The unit of analysis was each subsample (*unique combination of accommodation, test content, ELL status, and grade level*) that had a corresponding control group receiving the original test booklet version

Overall sample sizes for studies tended to be large but per subsample sizes varied widely



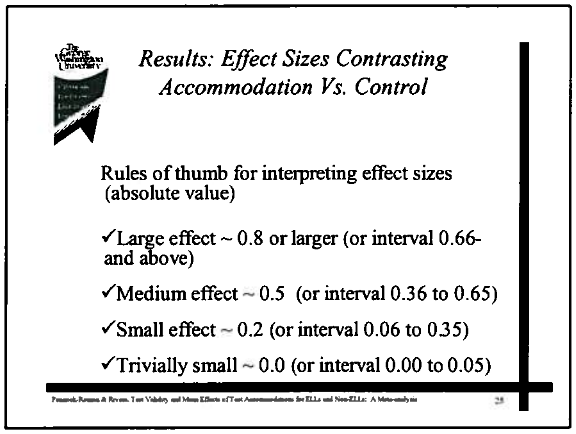
Results: Overall Descriptive Findings (cont.)

Mostly grade 8, math or science content

Items based mostly on NAEP and TIMSS items, not state assessments

44 subsamples of ELLs and 32 subsamples of Non-ELLs

Type of accommodations offered
- ✓ English language
  - 59% for ELLs, 73% for Non-ELLs
- ✓ Bilingual and native language
  - 34% for ELLs, 17% for Non-ELLs
- ✓ Extra time with no other changes
  - 5% for ELLs, 6% for Non-ELLs
- ✓ Small Group condition (one instance)
  - 2% for ELLs, 3% for non-ELLs



Results: Effect Sizes Contrasting Accommodation Vs. Control

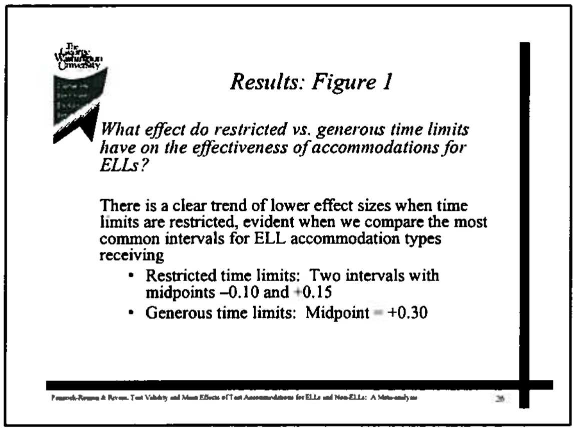
Rules of thumb for interpreting effect sizes (absolute value)

✓Large effect ~ 0.8 or larger (or interval 0.66- and above)

✓Medium effect ~ 0.5 (or interval 0.36 to 0.65)

✓Small effect ~ 0.2 (or interval 0.06 to 0.35)

✓Trivially small ~ 0.0 (or interval 0.00 to 0.05)

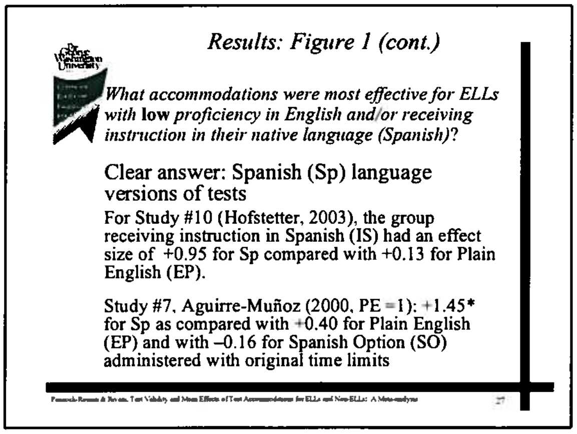


Results: Figure 1

What effect do restricted vs. generous time limits have on the effectiveness of accommodations for ELLs?

There is a clear trend of lower effect sizes when time limits are restricted, evident when we compare the most common intervals for ELL accommodation types receiving

- Restricted time limits: Two intervals with midpoints –0.10 and +0.15
- Generous time limits: Midpoint = +0.30



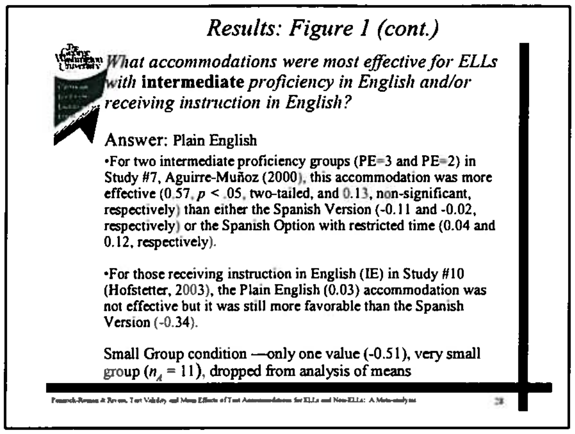
Results: Figure 1 (cont.)

What accommodations were most effective for ELLs with **low** *proficiency in English and/or receiving instruction in their native language (Spanish)?*

Clear answer: Spanish (Sp) language versions of tests

For Study #10 (Hofstetter, 2003), the group receiving instruction in Spanish (IS) had an effect size of +0.95 for Sp compared with +0.13 for Plain English (EP).

Study #7, Aguirre-Muñoz (2000, PE = 1): +1.45* for Sp as compared with +0.40 for Plain English (EP) and with –0.16 for Spanish Option (SO) administered with original time limits



Results: Figure 1 (cont.)

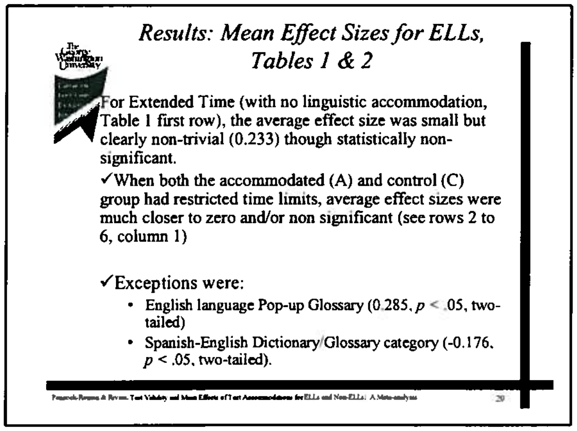
What accommodations were most effective for ELLs with **intermediate** *proficiency in English and/or receiving instruction in English?*

Answer: Plain English

•For two intermediate proficiency groups (PE=3 and PE=2) in Study #7, Aguirre-Muñoz (2000), this accommodation was more effective (0.57, p < .05, two-tailed, and 0.13, non-significant, respectively) than either the Spanish Version (-0.11 and -0.02, respectively) or the Spanish Option with restricted time (0.04 and 0.12, respectively).

•For those receiving instruction in English (IE) in Study #10 (Hofstetter, 2003), the Plain English (0.03) accommodation was not effective but it was still more favorable than the Spanish Version (-0.34).

Small Group condition —only one value (-0.51), very small group (n_A = 11), dropped from analysis of means



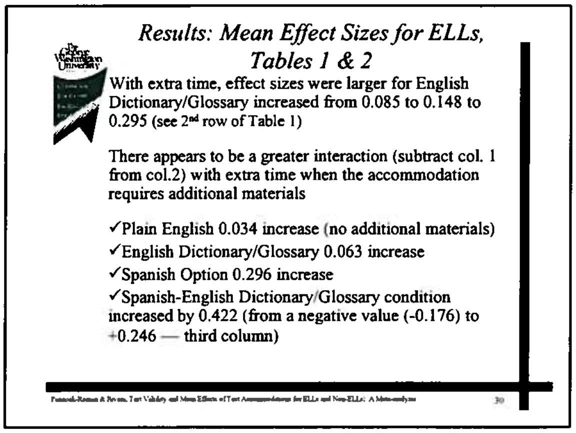
Results: Mean Effect Sizes for ELLs, Tables 1 & 2

For Extended Time (with no linguistic accommodation, Table 1 first row), the average effect size was small but clearly non-trivial (0.233) though statistically non-significant.

✓When both the accommodated (A) and control (C) group had restricted time limits, average effect sizes were much closer to zero and/or non significant (see rows 2 to 6, column 1)

✓Exceptions were:

- English language Pop-up Glossary (0.285, p < .05, two-tailed)
- Spanish-English Dictionary/Glossary category (-0.176, p < .05, two-tailed).



Results: Mean Effect Sizes for ELLs, Tables 1 & 2

With extra time, effect sizes were larger for English Dictionary/Glossary increased from 0.085 to 0.148 to 0.295 (see 2nd row of Table 1)

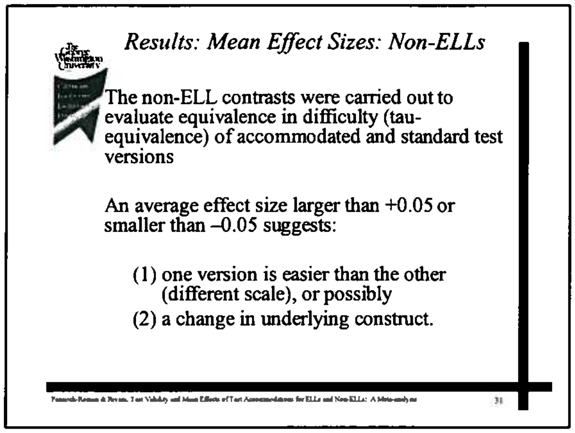
There appears to be a greater interaction (subtract col. 1 from col.2) with extra time when the accommodation requires additional materials

✓Plain English 0.034 increase (no additional materials)

✓English Dictionary/Glossary 0.063 increase

✓Spanish Option 0.296 increase

✓Spanish-English Dictionary/Glossary condition increased by 0.422 (from a negative value (-0.176) to +0.246 — third column)

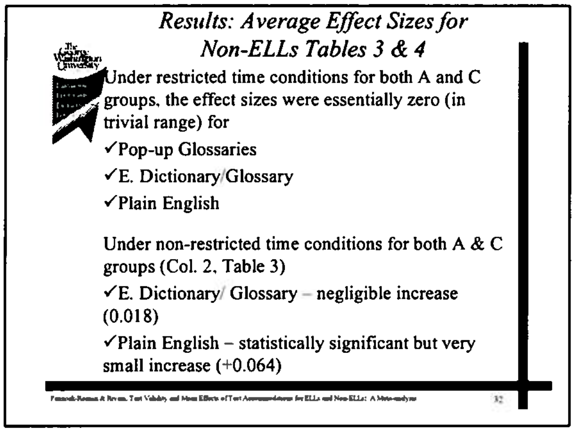


Results: Mean Effect Sizes: Non-ELLs

The non-ELL contrasts were carried out to evaluate equivalence in difficulty (tau-equivalence) of accommodated and standard test versions

An average effect size larger than +0.05 or smaller than –0.05 suggests:

(1) one version is easier than the other (different scale), or possibly
(2) a change in underlying construct.



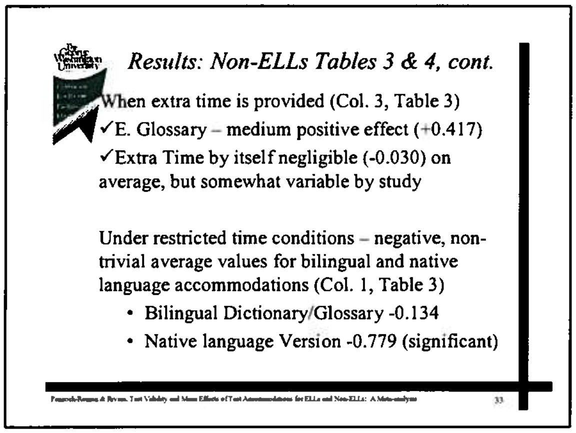
Results: Average Effect Sizes for Non-ELLs Tables 3 & 4

Under restricted time conditions for both A and C groups, the effect sizes were essentially zero (in trivial range) for

✓Pop-up Glossaries
✓E. Dictionary/Glossary
✓Plain English

Under non-restricted time conditions for both A & C groups (Col. 2, Table 3)

✓E. Dictionary/ Glossary – negligible increase (0.018)
✓Plain English – statistically significant but very small increase (+0.064)



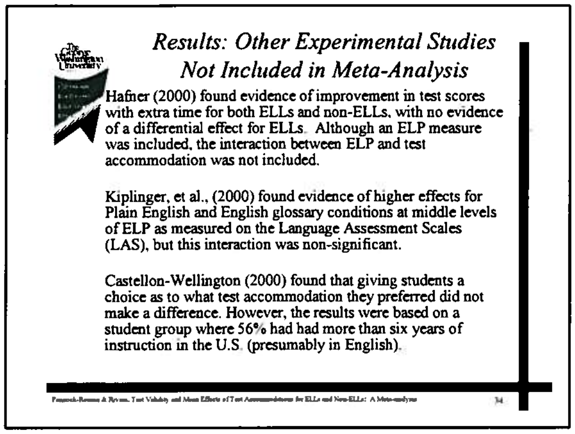
Results: Non-ELLs Tables 3 & 4, cont.

When extra time is provided (Col. 3, Table 3)

✓E. Glossary – medium positive effect (+0.417)
✓Extra Time by itself negligible (-0.030) on average, but somewhat variable by study

Under restricted time conditions – negative, non-trivial average values for bilingual and native language accommodations (Col. 1, Table 3)

- Bilingual Dictionary/Glossary -0.134
- Native language Version -0.779 (significant)

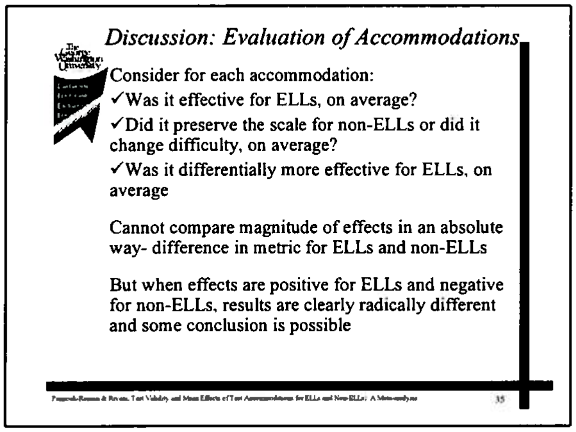


Results: Other Experimental Studies Not Included in Meta-Analysis

Hafner (2000) found evidence of improvement in test scores with extra time for both ELLs and non-ELLs, with no evidence of a differential effect for ELLs. Although an ELP measure was included, the interaction between ELP and test accommodation was not included.

Kiplinger, et al., (2000) found evidence of higher effects for Plain English and English glossary conditions at middle levels of ELP as measured on the Language Assessment Scales (LAS), but this interaction was non-significant.

Castellon-Wellington (2000) found that giving students a choice as to what test accommodation they preferred did not make a difference. However, the results were based on a student group where 56% had had more than six years of instruction in the U.S. (presumably in English).



Discussion: Evaluation of Accommodations

Consider for each accommodation:

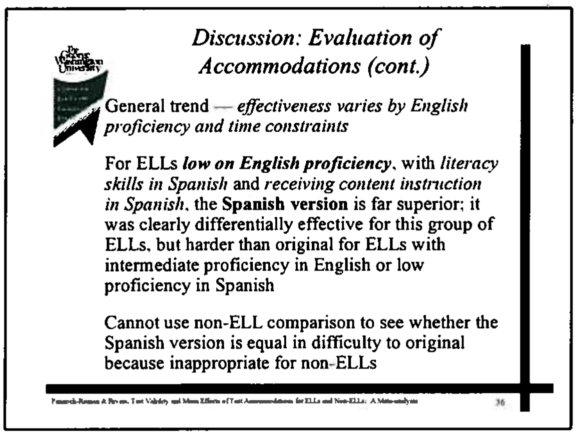
✓Was it effective for ELLs, on average?

✓Did it preserve the scale for non-ELLs or did it change difficulty, on average?

✓Was it differentially more effective for ELLs, on average

Cannot compare magnitude of effects in an absolute way- difference in metric for ELLs and non-ELLs

But when effects are positive for ELLs and negative for non-ELLs, results are clearly radically different and some conclusion is possible

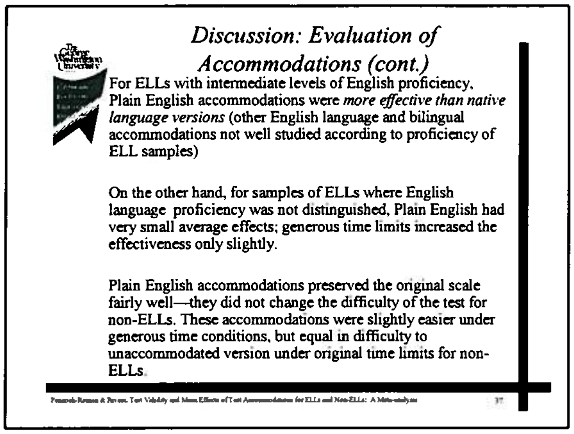


Discussion: Evaluation of Accommodations (cont.)

General trend — *effectiveness varies by English proficiency and time constraints*

For ELLs ***low on English proficiency***, with *literacy skills in Spanish* and *receiving content instruction in Spanish*, the **Spanish version** is far superior; it was clearly differentially effective for this group of ELLs, but harder than original for ELLs with intermediate proficiency in English or low proficiency in Spanish

Cannot use non-ELL comparison to see whether the Spanish version is equal in difficulty to original because inappropriate for non-ELLs

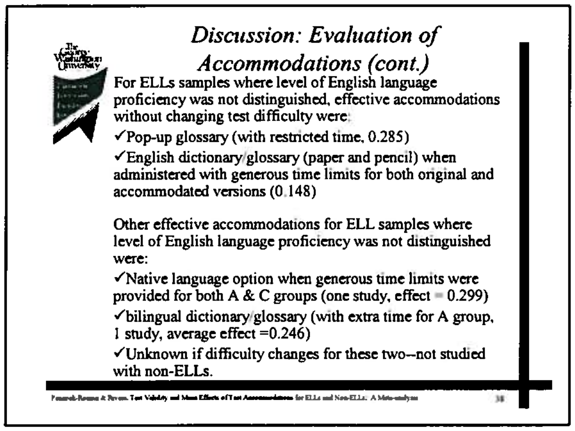


Discussion: Evaluation of Accommodations (cont.)

For ELLs with intermediate levels of English proficiency, Plain English accommodations were *more effective than native language versions* (other English language and bilingual accommodations not well studied according to proficiency of ELL samples)

On the other hand, for samples of ELLs where English language proficiency was not distinguished, Plain English had very small average effects; generous time limits increased the effectiveness only slightly.

Plain English accommodations preserved the original scale fairly well—they did not change the difficulty of the test for non-ELLs. These accommodations were slightly easier under generous time conditions, but equal in difficulty to unaccommodated version under original time limits for non-ELLs.



Discussion: Evaluation of Accommodations (cont.)

For ELLs samples where level of English language proficiency was not distinguished, effective accommodations without changing test difficulty were:

✓Pop-up glossary (with restricted time, 0.285)

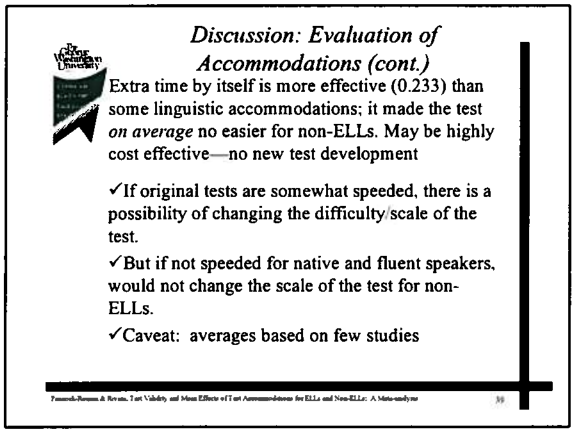
✓English dictionary/glossary (paper and pencil) when administered with generous time limits for both original and accommodated versions (0.148)

Other effective accommodations for ELL samples where level of English language proficiency was not distinguished were:

✓Native language option when generous time limits were provided for both A & C groups (one study, effect = 0.299)

✓bilingual dictionary/glossary (with extra time for A group, 1 study, average effect =0.246)

✓Unknown if difficulty changes for these two--not studied with non-ELLs.



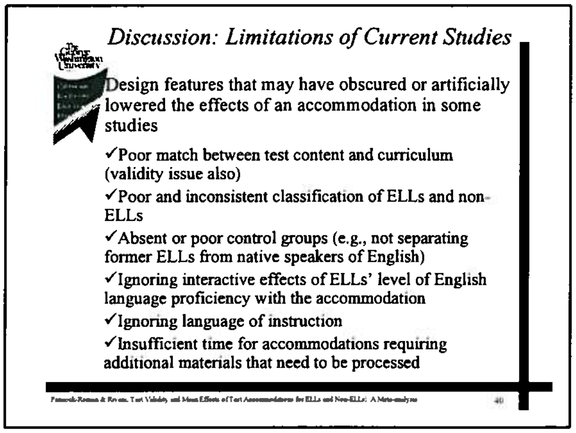
Discussion: Evaluation of Accommodations (cont.)

Extra time by itself is more effective (0.233) than some linguistic accommodations; it made the test *on average* no easier for non-ELLs. May be highly cost effective—no new test development

✓If original tests are somewhat speeded, there is a possibility of changing the difficulty/scale of the test.

✓But if not speeded for native and fluent speakers, would not change the scale of the test for non-ELLs.

✓Caveat: averages based on few studies



Discussion: Limitations of Current Studies

Design features that may have obscured or artificially lowered the effects of an accommodation in some studies

✓Poor match between test content and curriculum (validity issue also)

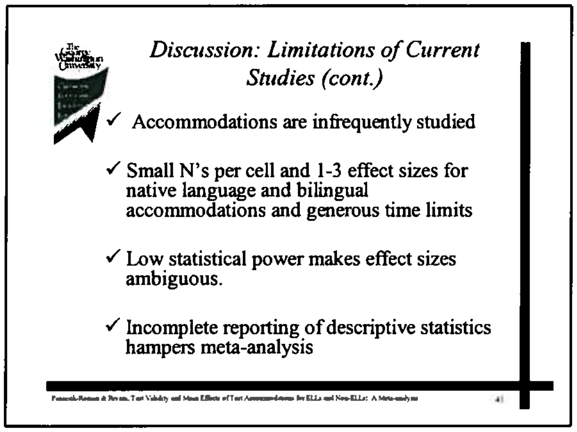
✓Poor and inconsistent classification of ELLs and non-ELLs

✓Absent or poor control groups (e.g., not separating former ELLs from native speakers of English)

✓Ignoring interactive effects of ELLs' level of English language proficiency with the accommodation

✓Ignoring language of instruction

✓Insufficient time for accommodations requiring additional materials that need to be processed



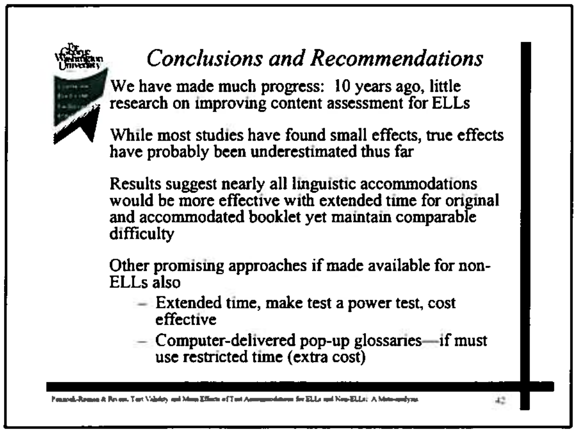
Discussion: Limitations of Current Studies (cont.)

✓ Accommodations are infrequently studied

✓ Small N's per cell and 1-3 effect sizes for native language and bilingual accommodations and generous time limits

✓ Low statistical power makes effect sizes ambiguous.

✓ Incomplete reporting of descriptive statistics hampers meta-analysis



Conclusions and Recommendations

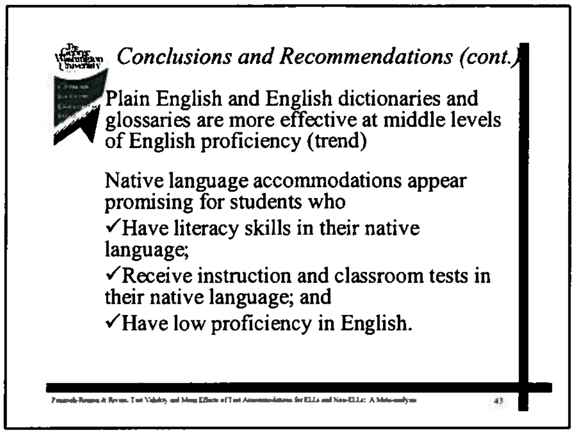
We have made much progress: 10 years ago, little research on improving content assessment for ELLs

While most studies have found small effects, true effects have probably been underestimated thus far

Results suggest nearly all linguistic accommodations would be more effective with extended time for original and accommodated booklet yet maintain comparable difficulty

Other promising approaches if made available for non-ELLs also

- Extended time, make test a power test, cost effective
- Computer-delivered pop-up glossaries—if must use restricted time (extra cost)



Conclusions and Recommendations (cont.)

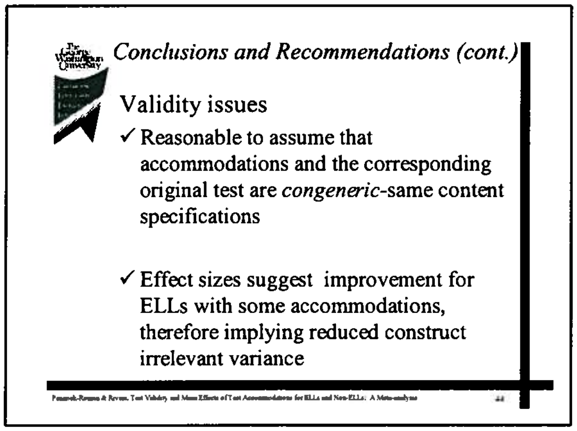
Plain English and English dictionaries and glossaries are more effective at middle levels of English proficiency (trend)

Native language accommodations appear promising for students who

✓Have literacy skills in their native language;

✓Receive instruction and classroom tests in their native language; and

✓Have low proficiency in English.

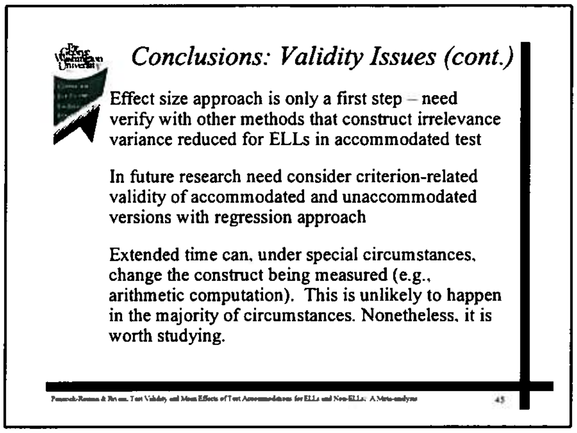


Conclusions and Recommendations (cont.)

Validity issues

✓ Reasonable to assume that accommodations and the corresponding original test are *congeneric*-same content specifications

✓ Effect sizes suggest improvement for ELLs with some accommodations, therefore implying reduced construct irrelevant variance



Conclusions: Validity Issues (cont.)

Effect size approach is only a first step – need verify with other methods that construct irrelevance variance reduced for ELLs in accommodated test

In future research need consider criterion-related validity of accommodated and unaccommodated versions with regression approach

Extended time can, under special circumstances, change the construct being measured (e.g., arithmetic computation). This is unlikely to happen in the majority of circumstances. Nonetheless, it is worth studying.

Thank You!

mpennock@mainc.rr.com crivera@ceee.gwu.edu

Pennock-Roman & Rivera. Test Validity and Mean Effects of Test Accommodations for ELLs and Non-ELLs: A Meta-analysis 46

Figure 1. Accommodated vs. Control Versions of Tests Administered to ELLs - Inside Out Display of Glass's Unbiased d Values

More Generous Time Limits or Untimed				Interval	More Restricted Time Limits			
Study, Acc, Grade/PE/IL, *d*	Study, Acc, Grade/PE/IL, *d*	Study, Acc, Grade/PE/IL, *d*	Study, Acc, Grade/PE/IL, *d*	Midpoint	Study, Acc, Grade/PE/IL, *d*	Study, Acc, Grade/PE/IL, *d*	Study, Acc, Grade/PE/IL, *d*	Study, Acc, Grade/PE/IL, *d*
				1.45	7, Sp, PE=1, 1.45			
				1.40				
				1.00 to 1.35				
				0.95	10, Sp, IS, 0.95			
				0.80 to 0.90				
				0.75				
			3, EP, 8, 0.71	0.70				
				0.65				
				0.60				
				0.55	7, EP, PE=3, 0.57			
			3, EDG, 4, 0.52	0.50				
			3, SEG, 4, 0.45	0.45	1, EGP, 4, 0.46			
				0.40	7, EP, PE=1, 0.40	6, EDG, 8, 0.41		
				0.35				
1, T, 4, 0.28	3, EDG, 8, 0.29	4, EDG, 8, 0.29	8, SO, 8, 0.30	0.30	1, EDG, 4, 0.29			
				0.25				
				0.20	11, EP, 4, 0.20	2, EP, 8, 0.20		
		3, EP, 4, 0.14	4, T, 8, 0.16	0.15	1, EDG, 8, 0.13	7, EP, PE=2, 0.13	10, EP, IS, 0.13	1, EGP, 8, 0.17
		12, EP, 8, 0.08	13, EDG, M, 0.08	0.10	2, EDG, 8, 0.10	4, EP, 8, 0.10	7, SO, PE=2, 0.12	
				0.05	6, SEG, 8, 0.03	10, EP, IE, 0.03	7, SO, PE=3, 0.04	
				0.00	7, Sp, PE=2, -0.02	2, EDG, 4, 0.01		
				-0.05	4, EDG, 8, -0.04			
				-0.10	2, SEG, 8, -0.12	5, EP, 8, -0.12	7, Sp, PE=3, -0.11	2, EP, 4, -0.09
				-0.15	7, SO, PE=1, -0.16			
				-0.20				
			3, SEG, 8, -0.23	-0.25				
				-0.30	2, SEG, 4, -0.28			
				-0.35	10, Sp, IE, -0.34			
				-0.40				
				-0.45				
				-0.50	1, GS, 4, -0.51			
				-0.55				
				-0.60				
				-0.65				
				-0.70	11, EP, 6, -0.66			
				-0.75				
				-0.80 to -0.90				
				-0.95				
				-1.00				
				-1.05				
				-1.10				
				-1.15	5, Sp, 8, -1.13			

Legend

Accomodations:

EDG = English Dictionary or Gloss

EGP = English Glossary, Pop-up

EP = English, Plain

GS = Groups, Small

SEG = Spanish-English Glossary

SO = Spanish Option/Dual Language

Sp = Spanish Version

T = Extra Time

Grade/Proficiency in English (PE)/

4, 6, or 8 = Grade

M = Middle School

PE = 1 (low) to 4 (high)

IE = Instruction in English

IS = Instruction in Spanish

Experimental Studies Listed in Figure 1[1]

Studies with Independent Groups Designs

1) Abedi, J., Courtney, M., & Leon, S. (2003a). *Research-supported accommodation for English language learners in NAEP (CSE Tech. Rep. No. 586).* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved December 1, 2004 from http://www.cse.ucla.edu/products/reports_set.htm

2) Abedi, J., Courtney, M., & Leon, S. (2003b). *Effectiveness and validity of accommodations for English language learners in large-scale assessments (CSE Technical Report No. 608).* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved December 1, 2004 from http://www.cse.ucla.edu/products/reports_set.htm.

3) Abedi, J., Courtney, M., Mirocha, J., Leon, S., and Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification* (CSE Report 666). Los Angeles: CA: National Center for Research on Evaluation, Standards, and Student Testing, University of California. Retrieved [9/28/06], from the World Wide Web: http://www.cse.ucla.edu/products/reports.asp

4) Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: Interactions with student language background* (CSE Tech. Rep. No. 536). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved December 1, 2004 from http://www.cse.ucla.edu/products/reports_set.htm.

5) Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Technical Report No. 478). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing. **(Adjusted descriptive statistics to remove overlap with Hofstetter, 2003)**

6) Abedi, J., Lord, C., Kim, C., & Miyoshi, J. (2001). *The effects of accommodations on the assessment of limited English proficient students in the National Assessment of Educational Progress* (Publication No. NCES 2001–13). Washington, DC: National Center for Education Statistics. Retrieved December 1, 2004 from http://nces.ed.gov/pubs2001/200113.pdf.

7) Aguirre-Muñoz, Z. (2000). *The impact of language proficiency on complex performance assessments: Examining linguistic accommodation strategies for English language learners*. Dissertation Abstracts International, A.(UMI No. 9973171).

8)Anderson, M., Liu, K., Swierzbin, B., Thurlow, M., & Bielinski, J. (2000). *Bilingual accommodations for limited English proficient students on statewide reading rests: Phase 2.* (Minnesota Rep. No. 31). Minneapolis, MN: National Center for Educational Outcomes. Retrieved December 1, 2004 from http://education.umn.edu/NCEO/OnlinePubs/MnReport31.html.

9) Duncan, T. G., Parent, L. del R., Chen, W., Ferrara, S., Johnson, E., Oppler, S., & Shieh, Y. (2005). Study of a dual language test booklet in eighth grade mathematics. *Applied Measurement in Education, 18*(2), 129–161.

10) Hofstetter, C. H. (2003). Contextual and mathematics accommodation test effects for English language learners. *Applied Measurement in Education, 16*(2), 159–188.

11) Rivera C., & Stansfield, C. W.(2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment, 9* (3 &4), 79-106.

Studies with Repeated Measures Designs

12*)* Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance.* (CSE Technical Report No. 429). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing. (Same data as in Abedi, J., & Lord, C. ,2001, The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.)

13) Albus, D., Bielinski, J., Thurlow, M., & Liu, K. (2001). *The effect of a simplified English language dictionary on a reading test.* (LEP Project Rep. No. 1). Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes. Retrieved December 1, 2004 from http://education.umn.edu/NCEO/OnlinePubs/LEP1.html.

[1] Numbered titles correspond to the studies that examined accommodations listed in Figure 1.

Table 1. Contrasting Average Values of *d* for ELLs for the Same Accommodation Across Time Conditions and Levels of Language Proficiency

Accommodations Subdivided by Time Constraints

	Both A & C Groups Receiving		Extended Time for A Group Only
	Restricted Time	Extended Time	
Expected Value			δ_T
Extended Time By Itself	N/A	N/A	0.233
Expected Value	δ_j	$\delta_j + \delta_{jxT}$	$\delta_j + \delta_{jxT} + \delta_T$
Pop-up Glossaries	0.285 *	not studied	not studied
English Dictionary/G.	0.085	0.148 *	0.295
Plain English	0.053	0.087 *	not studied
Spanish Option	0.003	0.299	not studied
Spanish-English D./G.	-0.176 *	not studied	0.246

Accommodations Subdivided by Language Proficiency**

	Low Proficieny-Spanish	Med Proficieny-English	Low Proficieny-English
Expected Value	δ_j	$\delta_j + \delta_{jxT}$	$\delta_j + \delta_{jxT}$
Spanish Version	-0.560 *	-0.072	1.323 *

* Significantly different from zero at $p < .05$ (two-tailed)

** In these subsamples,proficiency in one language is associated with low proficiency in the other.

Table 2. Mean Values for Effect Sizes for ELLs, Standard Errors, Confidence Intervals, and Components of Expected Values for 14 Categories of Accommodations

				95% C I					Components in Expected Values		
Category Number	Mean Effect Size		St Err	Lower Limit	Upper Limit	Subsamples	Total N (A)	Total N (C)	δ_i	δ_{iXT}	δ_T
Restricted Time Limits for A and C Groups											
1. Pop-up Glossaries	0.285	*	0.125	0.040	0.530	2	119	166	√		
2. English Dictionaries/G.	0.085		0.050	-0.013	0.183	6	827	835	√		
3. English, Plain	0.053		0.042	-0.029	0.136	11	1178	1183	√		
4. Spanish Option	0.003		0.125	-0.242	0.247	3	88	159	√		
5. Spanish-English D./G.	-0.176	*	0.067	-0.308	-0.045	3	324	525	√		
Both Groups Had Little or No Time Constraints											
6. Spanish Option	0.299		0.216	-0.123	0.722	1	53	52	√	√	
7. English Dictionaries/G.	0.148	*	0.064	0.022	0.274	3	215	217	√	√	
8. English, Plain	0.087	*	0.030	0.028	0.145	3	502	555	√	√	
A Group Extra Time, C Group No Extra Time											
9. English Dictionaries/G.	0.295		0.244	-0.183	0.772	1	29	144	√	√	√
10. Spanish-English D./G.	0.246		0.150	-0.048	0.540	2	80	84	√	√	√
11. Extra Time By Itself	0.233		0.131	-0.025	0.491	2	119	224			√
Spanish Versions Classified by Students' Level of Language Proficiency									δ_i	δ_{iXM}	δ_{iXH}
12. Low Spanish P.	-0.560	*	0.078	-0.712	-0.408	2	162	344	√		
13. Intermed. English P.	-0.072		0.164	-0.393	0.249	2	55	113	√	√	
14 .Low English P.	1.323	*	0.273	0.787	1.859	2	140	55	√		√

* Significantly different from zero at $p < .05$ (two-tailed)

Table 3. Contrasting Average Values of *d* for Non- ELLs for the Same Accommodation Across Time Conditions

Accommodations Subdivided by Time Constraints

	Both A & C Groups Receiving		Extended Time for A Group Only
	Restricted Time	Extended Time	
Expected Value			δ_T
Extended Time By Itself	N/A	N/A	-0.030
Expected Value	δ_j	$\delta_j + \delta_{jxT}$	$\delta_j + \delta_{jxT} + \delta_T$
Pop-up Glossaries	0.032	not studied	not studied
English Dictionary/G.	-0.004	0.018	0.417
Plain English	-0.008	0.064 *	not studied
Spanish Option/ Dual L.	-0.169	not studied	not studied
Spanish-English D./G.	-0.134	not studied	not studied
Spanish Version	-0.779 *	not studied	not studied

* Significantly different from zero at $p < .05$ (two-tailed)

Table 4. Mean Values for Effect Sizes for Non-ELLs, Standard Errors, Confidence Intervals, and Components of Expected Values for 10 Categories of Accommodations

				95% C I					Components in Expected Values		
Category Number	Mean Effect Size		St Err	Lower Limit	Upper Limit	# Sub samples	Total N (A)	Total N (C)	δ_j	δ_{jXT}	δ_T
Restricted Time Limits for A and C Groups											
1. Pop-up Glossaries	0.032		0.109	-0.180	0.245	2	112	229	√		
2. English Dictionaries/G.	-0.004		0.048	-0.098	0.090	6	871	924	√		
3. English, Plain	-0.008		0.017	-0.042	0.026	10	6571	6591	√		
4. Spanish Option	-0.169		0.138	-0.439	0.101	1	74	119	√		
5. Spanish-English D./G.	-0.134		0.070	-0.272	0.003	3	305	565	√		
Both Groups Had Little or No Time Constraints											
6. Spanish Option	Not administered to non-ELLs								√	√	
7. English Dictionaries/G.	0.018		0.079	-0.137	0.173	3	193	187	√	√	
8. English, Plain	0.064	*	0.028	0.010	0.118	2	569	631	√	√	
A Group Extra Time, C Group No Extra Time											
9. English Dictionaries/G.	0.417		0.216	-0.006	0.840	1	30	130	√	√	√
10. Spanish-English D./G.	Not administered to non-ELLs								√	√	√
11. Extra Time By Itself	-0.030		0.118	-0.261	0.201	2	109	228			√
Spanish Version									δ_j	δ_{jXM}	δ_{jXH}
12. Low Spanish P.	-0.779	*	0.182	-1.137	-0.422	1	24	61	√		

* Significantly different from zero at $p < .05$ (two-tailed)

Appendix 1: Technical Notes

Correction Factor for Individual Effect Sizes

Before calculation of the average effect sizes for each category, a correction factor was multiplied times each effect size, because the expected value of Glass's *d* index is known to overestimate its corresponding population value, particularly when the control group sample sizes are small (Hedges, 1981). The correction factor *c(v)* for degrees of freedom *v* given by Hedges is:

$$c(\nu) = \frac{\Gamma\left(\frac{\nu}{2}\right)}{\sqrt{\frac{\nu}{2}}\,\Gamma\left(\frac{\nu-1}{2}\right)},$$

where *Γ(v)* is the gamma function, which we calculated using Microsoft Excel. The values of this correction factor are always smaller than unity, and they approach unity as the sample size increases. They become smaller than unity by less than 0.0151 units once the sample size numbers above 50. Owing to the relatively low incidence of studies having fewer than 50 cases, the correction did not appreciably alter most indices.

Calculation of Average Effect Sizes for Categories of Accommodations

The calculation of each effect size average was based on values that could be considered statistically independent from others in the same average. The results at different grade levels within the same study were treated as separate, independent effect sizes because they were based on subsamples that comprised different individuals, each having its own separate control group. Although effect sizes for different conditions within the same study were statistically dependent if they shared the same control group, for each average there were no observations sharing the same control group. Hence, they can be considered statistically independent, except perhaps for any statistical dependence resulting from shared characteristics such as study authors, location, source of items, and timing. This possible dependence is a general limitation of these data given that many studies, even those completed at different times, share authors, the same pool of items, and similar geographical locations.

Weighting Schemes. For each of the accommodation categories having two or more independent effect sizes, the values corrected for bias were averaged using two different weighting schemes. One procedure weighted each effect size from the i^{th} study for the j^{th} accommodation type by the inverse of its sampling variance ($w_{ij} = 1/v_{ij}$). Another procedure weighted each effect size by the control group sample size $w_{ij} = n_{cij}$. Tables include only the mean based on inverse variance weights, which are considered superior because they minimize the variance of the average effect size (Shadish & Haddock, 1994). The mean using sample size weights was used only in the preliminary calculation of each effect size's sampling variance. As will be seen below, calculations for the sampling variances of individual effect sizes involve the population value for the mean effect size, which in turn depends on the inverse variance weights. Consequently, the estimate of the mean population value based on inverse variance weights was derived iteratively, with the individual variance calculations including the average based on sample size only in the initial iteration.[1] The general formula for calculating the weighted average effect size is given by (Shadish & Haddock, 1994, p. 265, equation 18.1):

$$\bar{d}_j = \frac{\sum_i w_{ij} d_{ij}}{\sum_i w_{ij}} \quad .$$

Sampling Variance of Individual Effect Sizes and Research Designs. For the first type of research design involving independent groups with possibly unequal variances, the normal approximation to the sampling variance of Glass's unbiased effect size using sample estimates of parameters is given by Gleser and Olkin (1994, 22-15, p.346):

$$v_{ij}^{U} = \frac{1}{n_j}\left[\frac{s_j^2}{s_C^2}\right] + \frac{1+\frac{1}{2}d_j^2}{n_C} \, ,$$

where vij is the sampling variance of the effect size for the ith study using the jth treatment type , nj and nC, Sj2 and SC2 are the sample sizes and variances for accommodated and standard test booklet groups, respectively, and dj is the jth accommodation average effect size (the estimate of the population average using the mean of observed effect sizes). Although the exact distribution for the effect size is a t distribution (Hedges, 1981), the normal approximation is very close to it with sample sizes of 30 or above and is a reasonable approximation for sample sizes of 10 or larger (Becker, 1988).

A different equation for the sampling variance is required when a single group receives both the accommodated and control booklets. In that case, the approximate sampling variance of the unbiased Glass index using sample estimates is given by Becker (1988, p. 263, Equation 13):

$$v_{ij}^{U} = \left[\frac{2(1-r_{ij})}{n_{ij}}\right] + \frac{d_{ij}^{\,2}}{2n_{ij}} \, ,$$

where d_{ij} is defined as before, r_{ij} is the correlation between accommodated and standard test scores,[2] and n_{ij} is the number of persons who took both the accommodated and standard test booklets.

Although combining effect sizes from studies having different designs can sometimes introduce some systematic biases into the average effect size (Morris & DeShon, 2002), the studies included here have the design features that Morris and DeShon recommended in order to minimize such biases. That is, the possible bias due to selectivity of the samples in the independent group design is minimized here because individuals were randomly assigned to accommodated and control conditions. Also, the possible bias in effect size due to a practice effect or changes over time in the repeated measures design found by Morris and Deshon is minimized here because the test versions were administered in a counter-balanced order.Nevertheless, before combining effect sizes from different research designs within the same category, we examined whether the studies with repeated-measures designs had systematically higher effect size estimates than did other studies, as found by Morris and Deshon. The values for Abedi, et al. (1997) and Albus, et al. (2001) were actually in the middle of the range of values for their corresponding category (see Results below, Table 1 and Figure 2). Therefore all subsamples for the same accommodation category were grouped together regardless of design.

***Sampling Variance of* Averages *for Effect Size*.** Under a fixed-effects model,[3] the sampling variance of the mean effect size for the j^{th} accommodation weights is given by (Shadish & Haddock, 1994, p. 266):

$$v_j = \frac{1}{\sum_i w_{ij}},$$

where $w_{ij} = 1/v_{ij}$ when using inverse variance weights and $w_{ij} = n_{iC}$ when the control group sample sizes are used as weights. The standard error of estimate for the average effect size j is the square root of this sampling variance $(v_j)^{½}$. A 95% confidence band can be calculated around the average result d_j using the unit normal value ($z = 1.96$) at $\alpha = .05$:

$$d_j \underline{+} 1.96(v_j)^{½}.$$

Heterogeneity of Effect Sizes within a Category. In order to test the hypothesis that all studies included in the calculation of a particular average (i.e., categorized in the accommodation type) have the same population effect size, we use the Q statistic (Shadish & Haddock, 1994) which takes the form:

$$Q = \sum_i \left[\frac{\left(d_{ij} - \bar{d}_j\right)^2}{v_{ij}} \right].$$

The terms above are defined as before, $i = 1$ *to* k, and the Q statistic is distributed as a chi-square with k-1 degrees of freedom.

Endnotes

[1] The subsequent new estimate of the average value using inverse sample weights substituted the mean based on sample sizes in the second and later iterations. Iterations continued until the change in the average based on inverse variances was negligible.

[2] Unfortunately, the authors of the studies using repeated measures designs did not report the needed correlations. Nevertheless, we were able to derive them for the combined ELL and non-ELL groups using the test statistics for repeated measures (t and F) together with other descriptive statistics reported by the authors.

[3] Owing to the small number of studies we were able to find, which can be considered the universe of scientific interest, we opted for a fixed effects approach instead of a random effects approach. As stated by Raudenbush (1994), a random effects model is preferred when the "studies under synthesis can be viewed as representative of a larger population or universe of implementations of a treatment " (p. 316) [to which we want to generalize]. "A conception of study effects as fixed versus random depends in part on the number of studies available" (p. 307). He points out that generalizing to a universe of studies appears ludicrous when there are few studies.