

Checking the Believability of Equating Results

Richard Hill
Center for Assessment
CCSSO NCSA Conference
Orlando, FL—June 17, 2008

Last Year

- **Requirements for Comparability**
 - Equivalent test
 - Equivalent population
 - Equivalent conditions
 - Equivalent scoring
- **Need to start at an earlier point—Do I believe the equating results?**

Solid Equating

- All changes in total results are based *entirely* on changes in the equating items
- Were equating items presented identically?
 - Same item—no alterations
 - Same equivalent position
- Were equating items representative?

Solid Equating (cont'd)

- **Was equating sample representative?**
- **Were any equating items deleted during the post-administration analysis?**

Observations

- All equating items in same position
- Approximately equal numbers of students taking each form, and modestly increasing N across years
- Mean change in p-values small
- Standard deviation of p-value change small
- Conclusion—Changes in mean scaled scores **MUST** be small and positive

When Things Aren't So Straightforward

- **Concern of state testing director that scores had increased too much**
- **Most extreme—Grade 5 reading**
 - Mean scaled score increase from 278 to 295 (standard deviation of 72)
 - Percent Proficient from 39 to 52

Table 2

- Equating items moved (but balanced)
- Some significant differences in N
- Most importantly, vastly different changes in p-values
- First, don't take an average

Summary of Table 2

Set	2006			2007			Δ Pos	ΔP
	Form	Ave p	N	Form	Ave p	N		
1	3	68	4087	C	71	32965	-7	+3
2	8	68	3965	1	59	4158	+7	-9
3	8	68	3965	2	67	3996	+7	-1
4	7	65	4089	3	68	3980	0	+3
5	7	65	4089	4	68	3980	0	+3

Standard Error

$$SE_{p_2 - p_1} = \sqrt{\frac{p_1 * (1 - p_1)}{N_1} + \frac{p_2 * (1 - p_2)}{N_2}}$$

- About 1 percentage point
- Therefore, these results are way beyond chance

Figuring It Out

- **Step 1—Increased N for Form 1**
- **Forms were not uniformly distributed—Form 1 held out for use with special populations**

Deviation from State Average on Common Items

Form	2006	2007
1	-4.5	-9.1
2	-0.2	+0.2
3	+0.9	+2.3
4	+1.6	+1.9
5	+0.6	+1.4
6	+1.8	+2.0
7	+0.7	+1.8
8	-0.6	+1.3

Changes on Equating Items After Adjustment

2006 Form	2007 Form	Δp	2006 Adj.	2007 Adj.	Net
3	C	+3.0	+0.9	0.0	+3.9
8	1	-8.5	-0.6	+9.1	0.0
8	2	-0.5	-0.6	-0.2	-1.3
7	3	+3.7	+0.7	-2.3	+2.1
7	4	+3.8	+0.7	-1.9	+2.6

Adjusted Change in p by Change in Item Location

Pos. Change	Grade				
	4	5	6	7	8
+7	-2.8	-0.7	-3.9	N/A	-3.4
0	-0.5	+2.3	+1.3	+1.0	+0.6
-7	+2.3	+3.9	+2.9	+3.8	+2.2

Effect of Position Change for Reading

- **0.4 percentage points per position**
- **Finding replicated in another state**
- **Finding did NOT apply to math**

Reasonableness of Reported Gains

- **Should 2 percentage points produce mean SS gain of 18?**
- **Effect size**
 - SD of p-values ~ 20
 - SD of SS ~ 70
- **No—p-values show effect size of .1; SS show effect size of .2**

Final Results

- Different equating model showed gain of 11 SS points
- No counterbalancing of items at grades 3 and 7 (which showed most gain after grade 5), so those results are suspect

Other Issues

■ **Constructed-response items**

- Must rescore last year's papers in this year's scoring process
- Working on process to take changes in difficulty of scoring into account

■ **Comparing PACs**

- Often considered the bottom line
- Granularity can lead to misleading results, so when checking equating, use mean SS

Recommendations

- **Collect and review data in Table 1 for every subject and grade**
- **Look at changes in average p-values (if data suggest it makes sense to take an average)**
- **Question any average change of more than 2 percentage points or SD greater than 1**

Questioning Unrealistic Changes in p-values

- **Look at last year's paper**
 - Differential content?
 - Unchanged location?
 - Identical presentation?
 - Unchanged administration procedures?
 - Tested populations equivalent?
- **Equating cannot fix any of these problems (GIGO)**

Changes in SS Consistent with Changes in p ?

- **Look at consistencies in effect size**
- **Equating should not be accepted as correct until inconsistencies can be explained**

Copies of Paper

- www.nciea.org