How Insensitivity to Instruction Relates to Reduced Validity and Distorted Accountability
What has happened in two years?

From new $450 M+ contract to ...

... what a reporter for Dallas Morning News calls a “hermeneutics of skepticism”

“What is the evidence, at scale, that the tests are sufficiently sensitive to instruction to warrant their continued use as part of a high stakes accountability system?”

HB 5 passed nearly unanimously & signed [# HS test from 15 to 5]

HB 2836 passed unanimously but vetoed
[ independent analysis of reliability and validity ...]

So can sensitivity implicate validity
Has Item Response Theory Increased the Validity of Achievement Test Scores?

Robert L. Linn

Center for Research on Evaluation, Standards, and Student Testing
University of Colorado at Boulder

It is clear that item response theory (IRT) has been the basis of many important contributions to educational measurement and has opened up a number of new possibilities such as adaptive and customized testing. The focus of this article, however, is on the question of the contribution of IRT to the validity of interpretations of achievement test results in the context of four specific applications: construction of scales for achievement tests; test construction; development of customized tests; and investigation of the influence of instruction on achievement tests. It is concluded that content considerations deserve greater attention in applications of IRT with achievement tests than they seem to have been given in these contexts.

Insensitivity Can Implicate Validity [of inferences] (of psychometric model – 1 PL)

“Fit” of Model
What kinds of analyses could address sensitivity-related validity

Test 1 :: Analyses of “What works ...”  
(Successful interventions)

Test 2 :: Equivalence Class Analysis

Test 3 :: Variance Associated with Membership in Score-Based equivalence classes

Test 4 :: How Like/Un-Alike are Item Response Functions?

What would/does this mean?

← 100% of Variance →

[side note – why not done?]

[side note re: DIF]

[side note re: Face Validity ]
Test 1

Analyses of “What works ...”
(Successful interventions)
<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>2006 TNCE Score</td>
<td>0.706</td>
<td>0.030</td>
<td>0.759</td>
<td>23.560</td>
</tr>
<tr>
<td>Study Treatment</td>
<td><strong>5.263</strong></td>
<td><strong>1.166</strong></td>
<td><strong>0.143</strong></td>
<td><strong>4.512</strong></td>
</tr>
<tr>
<td>Minority</td>
<td>-3.450</td>
<td>0.744</td>
<td>-0.110</td>
<td>-4.639</td>
</tr>
</tbody>
</table>

Dependent variable – 2007 TNCE

$R^2 = 0.504$, Durbin-Watson = 1.970, N = 1030

~ 5 point increase; 14% of Variance

(OLS: ~ 6.5 point increase; 17% of Variance)
<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>2006 TNCE Score</td>
<td>0.706</td>
<td>0.030</td>
<td>0.759</td>
<td>23.560</td>
</tr>
<tr>
<td>Study Treatment</td>
<td>5.263</td>
<td>1.166</td>
<td>0.143</td>
<td>4.512</td>
</tr>
<tr>
<td>Minority</td>
<td>-3.450</td>
<td>0.744</td>
<td>-0.110</td>
<td>-4.639</td>
</tr>
</tbody>
</table>

Dependent variable – 2007 TNCE

$R^2 = 0.504$, Durbin-Watson = 1.970, N = 1030
Two Quick “So What?” Responses:

Might not be good intervention(s)

Insensitivity has been around forever ...  
Linn: If valid before, then IRT could make more valid(robust etc.) ... done.
EQUVALENCCE CLASS ANALYSIS
Why Equivalence Class? – Minimal Account of Scoring (no additional assumptions) 
Allows us to examine pairing of scores for members of equivalence classes

Three Examples of Pairings by Equivalence Classes
[1] Gender Equivalence Classes – Male/Female 

\[
\begin{align*}
M1 & \leftrightarrow S_{M1} \\
M2 & \leftrightarrow S_{M2} \\
M3 & \leftrightarrow S_{M3} \\
M4 & \leftrightarrow S_{M4} \\
M5 & \leftrightarrow S_{M5} \\
\end{align*}
\]

Scores Assigned

Males \leftrightarrow Mean S_{\text{Male}}

\[
\begin{align*}
F1 & \leftrightarrow S_{F1} \\
F2 & \leftrightarrow S_{F2} \\
F3 & \leftrightarrow S_{F3} \\
F4 & \leftrightarrow S_{F4} \\
F5 & \leftrightarrow S_{F5} \\
\end{align*}
\]

Scores Assigned

Females \leftrightarrow Mean S_{\text{Female}}

COMPARISON OF PAIRING FOR TWO EQUIVALENCE CLASSES
Histogram of students’ 4th grade math scores in 2004 by item
Histogram of students’ 4th grade math scores in 2004 by item
Histogram of students’ 4th grade math scores in 2004 by item
VARIANCE ASSOCIATED WITH MEMBERSHIP IN SCORE-BASED EQUIVALENCE CLASSES
SPLIT THE DATA SET FOR TEXAS

Use ½ the Data Set To Characterize the Items ($b_1, b_2, b_3, b_4 \ldots$ for each item)

We’ll use these values to explore how much of the variance in this year’s scores is accounted for based on treating the previous year’s scores a set of equivalence classes and as measure of student “ability”.

$\frac{1}{2}$ to compare to $b$-based expected values for students in respective score-based equivalence classes

HOW MUCH OF THE VARIANCE ACCOUNTED FOR? (Independent of Instruction)

(Equivalence Classes of Students Receiving Same Scores (‘Ability’) in Previous Years)

$S_1, S_2, S_3, S_4, S_5 \ldots$$E_1, E_2, E_3, E_4, E_5 \ldots$

Actual Scores

Expected Value/Scores
Use $\frac{1}{2}$ the Data Set To Characterize the Items ($b_1, b_2, b_3, b_4 \ldots$ for each item)

We’ll use these values to explore how much of the variance in this year’s scores is accounted for based on treating the previous year’s scores a set of equivalence classes and as measure of student “ability”.

Use $\frac{1}{2}$ to compare to b-based expected values for students in respective score-based equivalence classes

65-75% of Variance
Convergence – Variance (at Scale) that is Insensitive to Instruction

Empirical Results from Successful Interventions (i.e., where instruction is varied)

65-75% of Variance Insensitive to Instruction

Simulated Results based on Large-N Properties of the Items
Convergence – Variance (at Scale) that is Insensitive to Instruction

Empirical Results from Successful Interventions (i.e., where instruction is varied)

65-75% of Variance Insensitive to Instruction

Simulated Results based on Large-N Properties of the Items

Prior Knowledge?
(e.g., In math knowing multiplication tables)
Convergence – Variance (at Scale) that is Insensitive to Instruction

65-75% of Variance Insensitive to Instruction

Empirical Results from Successful Interventions (i.e., where instruction is varied)

Simulated Results based on Large-N Properties of the Items

Doesn’t behave like prior knowledge in the domain … E.g., knowing multiplication tables shouldn’t show up in English

PRIOR KNOWLEDGE? (e.g., in math, knowing multiplication tables)
Test 4

HOW LIKE/UNALIKE ARE ITEM RESPONSE FUNCTIONS
Logistic Curves for Two Items

Higher ability means more likely to get the item correct.

Harder items require higher ability to have the same likelihood of getting correct.
Ideally, a logistic curve for one subject area item should be not be related to the logistics curves for items (with the same Theta) in other subject areas. [If criterion referenced is to have any meaning, something like this would have to be the case].

This independence is simulated (to show what it should look like) above ... note, there is no relationship between the red curve and the arrangement of the points in the other subject areas. [R-Reading, M-Mathematics, H-History, S-Science; the numbers refer to specific items (to be) compared].
Compare Items Across Domains (from different scales) with b-value of about -0.90

What happens TAKS 2004
And this ...
And this ... in THREE SEPARATE YEARS of test administration.
65-75% of Variance
Uncharacterized Latent Trait

If not domain specific, then what is it? (see Pham for SEM)

“Test Taking Ability” (because that’s what the students are doing

“Flaw” in the Psychometric (IRT) Model - Fails to Fit the Phenomena) – Not Valid

Insensitivity to Instruction at the Item Level Does Implicate Validity

~20% of Variance
Sensitive to DIFFERENCES IN INSTRUCTION (within a given academic year)

Need to develop for “Down-the-Hall” Sensitivity to Differences in Instruction
If performance is evaluated in terms of improving scores, temptation will be to TARGET the Uncharacterized Latent Trait (or Test Taking Ability)

Test Prep becomes over emphasized (not low level test taking skills but more along the lines of [old] SAT prep ... the parallels also include the language of “college readiness” [sounds more like a description of what the SAT is/was about ... ]

Not at all clear that value-added approaches will/can change this focus.
And finally, when tests are developed in ways that preserve rank ordering the implications for any meaningful notion of equity or fairness are not what the NCLB or RTTT were to be about.

(Tests that preserve rank order produce graphs of results that are “locally parallel” [go up and down together] ... exactly like they do in this graph of TAKS results).
end
Computer Modeling of the Instructionally Insensitive Nature of the Texas Assessment of Knowledge and Skills (TAKS) Exam

by

Vinh Huy Pham, Bachelor's of Science in Biology and Chemistry

Dissertation
Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy