

Technical Considerations in Marketbasket Reporting

Robert J. Mislevy
Educational Testing Service
February 7, 2000

Presented at the National Research Council's Workshop on Marketbasket Reporting for
NAEP, February 7-8, 2000, Washington, D.C.

Overview

Why do people think they'd like marketbasket reporting?

Some disturbing truths

The organizations of the paper

- Data collection methods
- Targets of inference
- Reporting Scales
- Some inferential tools
- The matrix

Some examples

Conclusion

The Appeal of Marketbasket Reporting, Part I

In most domains of knowledge, we develop very powerful theories when we are very young. ...

No one has to tell a kid that heavy objects fall more quickly than light objects. It's totally intuitive. It happens to be wrong. ...

Experts are people who actually think about the world in more sophisticated and different kinds of ways. ... In your area of expertise, you don't think about what you do as you would when you were five years of age. But I venture to say that if I get to questioning you about something that you are not an expert in, the answers you give will be the answers you would have given before you had gone to school.

Howard Gardner, 1993, p. 5.

The Appeal of Marketbasket Reporting, Part II

- Quantum mechanics
 - At a level too small for us to see, things don't work at all like “intuitive physics” tells us.
 - You have to work with probability distributions for entities, not discrete points.
 - The probability distribution for one entity depends on other entities and their situations in nonintuitive, disturbing, ways.
- My car radio
 - I have no idea how it actually works.
 - It has an interface that supports a user model I can understand, and use to make the radio do what I want.

The Appeal of Marketbasket Reporting, Part III

It offers an intuitive interface for understanding assessment results.

YES.

It eliminates complicated statistical methods from assessment design and analysis.

In general, NO.

Some Disturbing Truths

- Except in very special cases, there is no single-point, examinee-at-a-time matchup between the evidence for a given inference about an examinee's proficiency from different test forms.
- In general, to get the right answer you have to work with entire probability distributions that capture both what you know about each examinee's proficiency and what you don't know.
- The proportion of examinees with observed scores above a given level is generally not a good estimate of the proportion of examinees above that point. (Observed score distributions are not the same as true score distributions, and they depend on the particulars of the test form.)
- In inferences involving ensembles and background variables, the probability distributions associated with a given examinee depend on her background data and the response and background data of other examinees.

Organization of the Paper

A matrix with ...

- Columns: Data collection methods
- Rows: Reporting scales
- Cells: How to estimate features of population and subpopulation distributions on the designated reporting scale.

Data Collection Methods

- Single “administerable” form, everybody takes it.
- Parallel administerable forms, everybody takes one.
- Tau-equivalent forms: Same structure, but different lengths. Percents-correct have same expectation. Different reliabilities.
- Congeneric forms: Same skills and item-types, but differ as to difficulty, length, and accuracy at different parts of the scale. Eg: hard forms & easy forms; CAT.
- Arbitrary forms, with respect to content and/or length. Current NAEP. Can’t assume forms can be fit with a single unidimensional scale.

Targets of Inference

- Overall and subgroup means
- Overall and subgroup standard deviations
- Overall and subgroup proportions of students above cut-points (PACs).

Eg: Proportions of students above NAEP achievement levels.

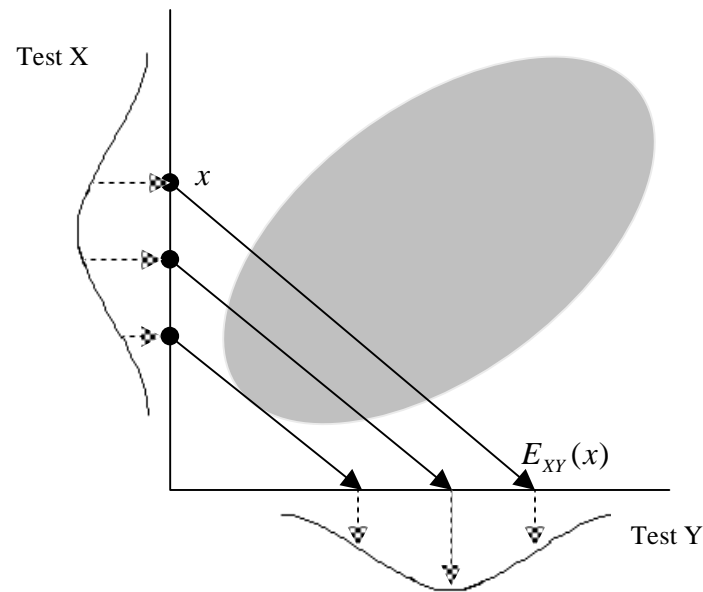
Reporting Scales

- MB1: Observed score on administerable form. One special form; observed scores on it is the reporting metric.
- MB2: True-score on administerable form. One special form; true scores on it is the reporting metric.
- MB3: True-score on nonadministerable form. Synthetic large form, to ensure content coverage. *True scores* that would, hypothetically, apply to this test form.
- MB4: Observed score on nonadministerable form. Synthetic large form. *Observed scores* that would apply to this test form.
- “Observed-scores,” or point estimates for individual students, on latent-variable scale(s). Issue: Even if IRT model is true, point estimates generally have different distributions for different test forms.
- “True-score” on latent-variable scale(s). Current NAEP.

Ten Slides on Tools

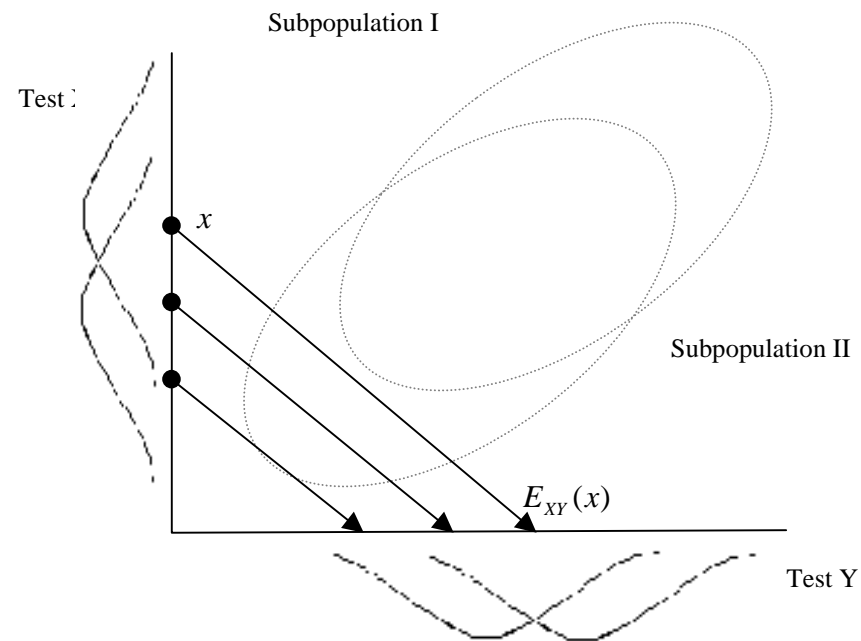
- Equating
 - Intuitive Test Theory
 - Presumes 1-to-1 matchup of *evidentiary value*
- One-stage empirical projection
- One-stage latent-variable projection
 - Without and with background variables
- Two-stage projection

Equating: A Single Population



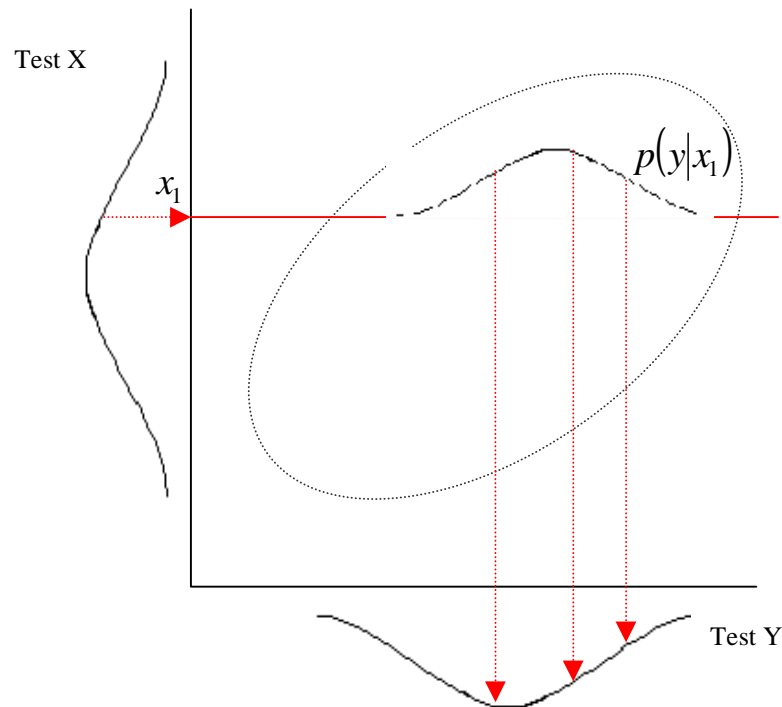
Equating. The link from the distribution of X scores to the distribution of Y scores is direct, and point-by-point. The chain of reasoning does not depend directly on $p(x,y)$, the joint distribution of X and Y (although $p(x,y)$ may be examined to support or weaken the linking argument). This kind of link must be justified by an extra-statistical argument; e.g., random sampling from the same pool, or parallel construction of matched items.

Equating, with Background Variables



Equating and subpopulations. When the requirements for equating are satisfied, the same point-by-point, examinee-by-examinee, linking holds regardless of background variables.

One-stage empirical projection, Without background variables

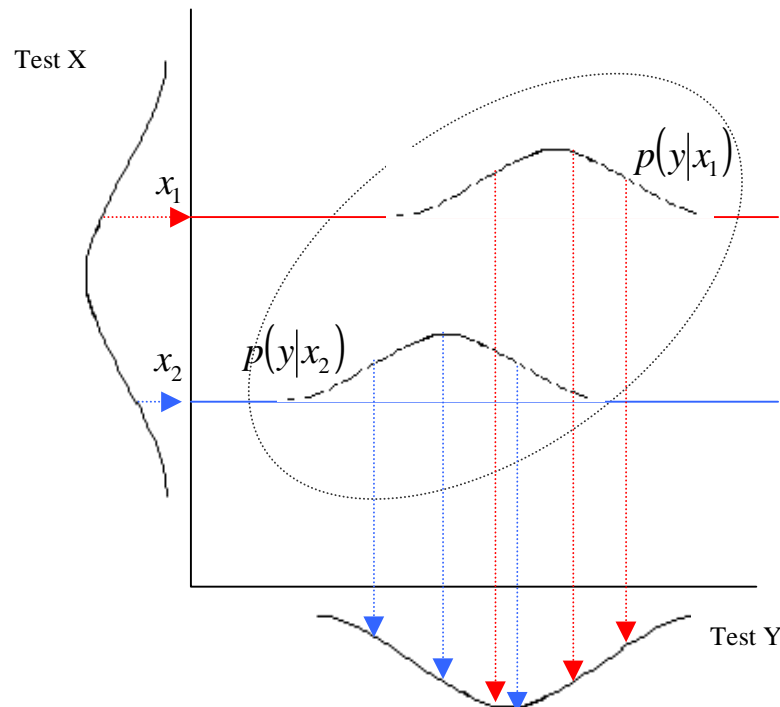


One-stage empirical projection. The link from the distribution of X scores to the distribution of Y scores is obtained by summing over students the predictive distributions:

$$p(y)|\mathbf{X} \approx N^{-1} \sum p(y|x_i).$$

The joint distribution $p(x,y)$ that the predictive distributions are based on is estimated from data, and may be mediated through a model, but it is not a latent-variable model.

One-stage empirical projection, Without background variables

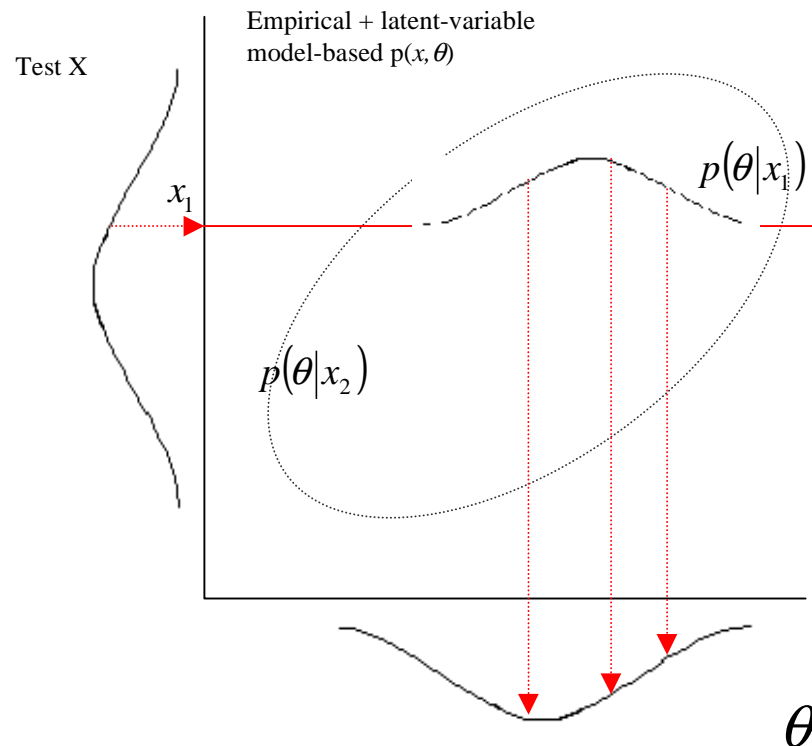


One-stage empirical projection. The link from the distribution of X scores to the distribution of Y scores is obtained by summing over students the predictive distributions:

$$p(y)|\mathbf{X} \approx N^{-1} \sum p(y|x_i).$$

The joint distribution $p(x,y)$ that the predictive distributions are based on is estimated from data, and may be mediated through a model, but it is not a latent-variable model.

One-stage latent-variable projection

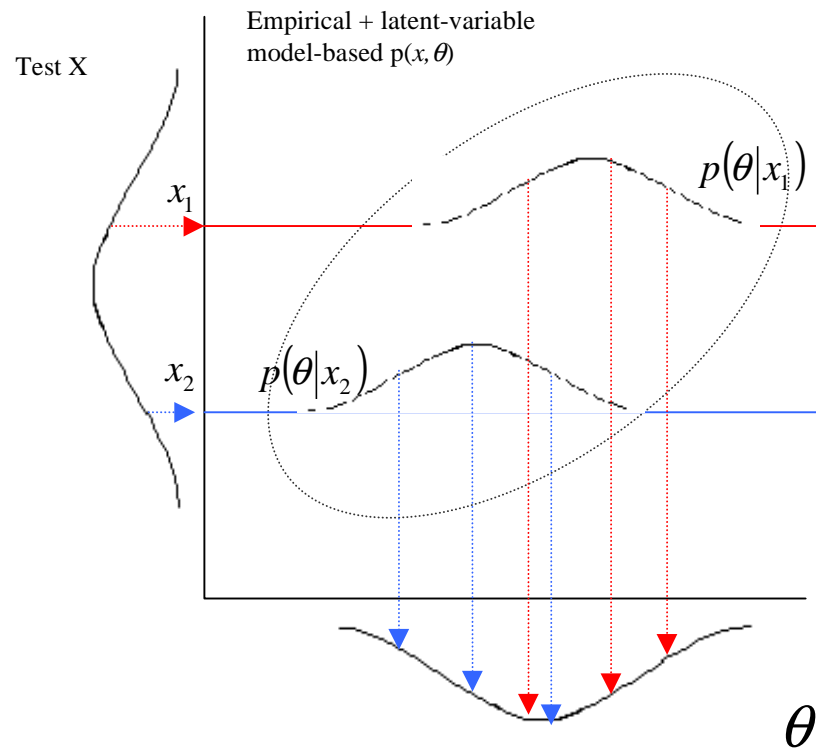


One-stage latent-model based projection. The link from the distribution of X scores to the distribution of θ values is obtained by summing over students the predictive distributions:

$$p(\theta)|\mathbf{X} \approx N^{-1} \sum p(\theta|x_i).$$

The joint distribution $p(x, \theta)$ on which the predictive distributions are based is empirical as far as x is concerned, but $p(x|\theta)$ is given by the latent-variable model.

One-stage latent-variable projection

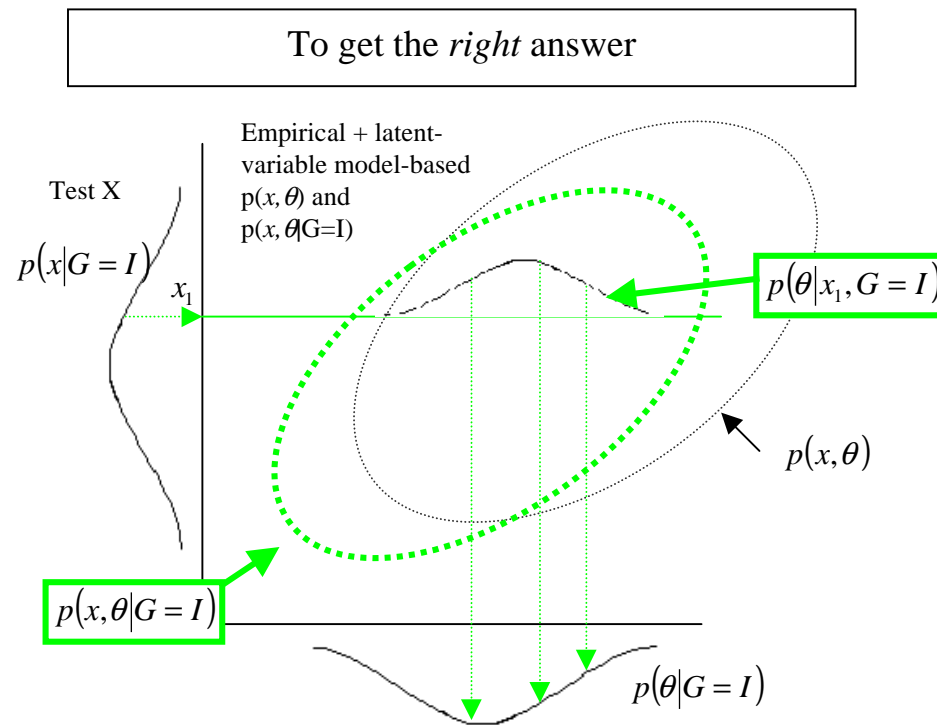


One-stage latent-model based projection. The link from the distribution of X scores to the distribution of θ values is obtained by summing over students the predictive distributions:

$$p(\theta)|\mathbf{X} \approx N^{-1} \sum p(\theta|x_i).$$

The joint distribution $p(x, \theta)$ on which the predictive distributions are based is empirical as far as x is concerned, but $p(x|\theta)$ is given by the latent-variable model.

One-stage latent-variable projection, with Background Variables

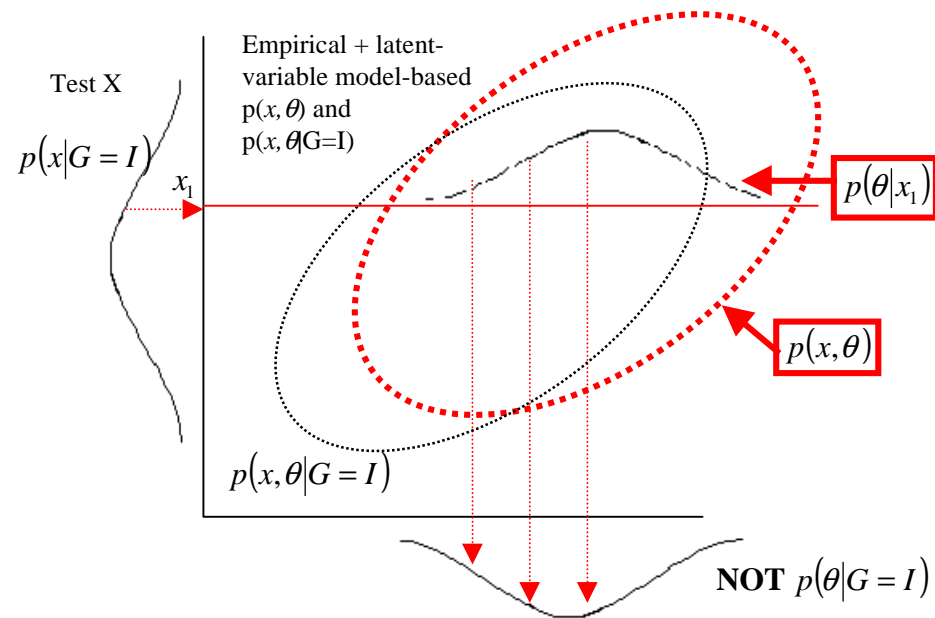


One-stage latent-model based projection, with background variables. The link from the distribution of X scores to the distribution of θ values for, say, Group= I , is obtained by summing over students the corresponding conditional predictive distributions:

$$p(\theta|G=I)|\mathbf{X} \approx N^{-1} \sum_i p(\theta|x_i, G=I).$$

One-stage latent-variable projection, with Background Variables

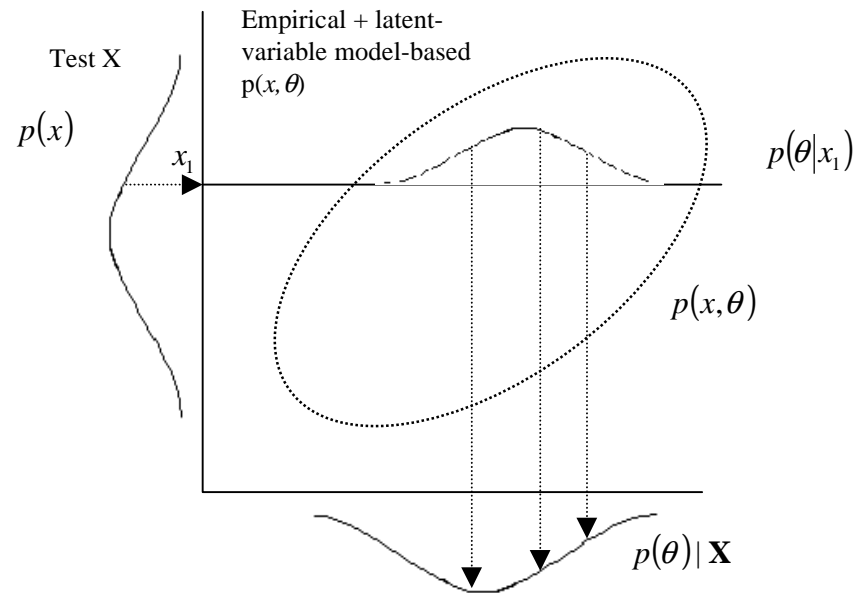
To get the *wrong* answer



The wrong answer is obtained if the predictive distribution does not include the covariates; that is,
$$p(\theta|G=I)|\mathbf{X} \neq N^{-1} \sum_{i:G=I} p(\theta|x_i).$$

Two-stage projection

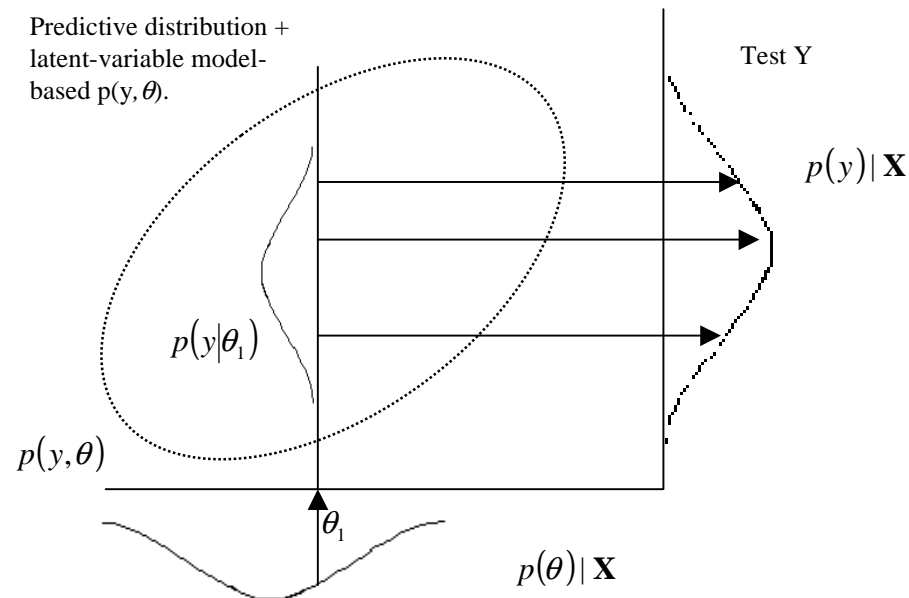
Stage 1: Predictive Distribution for the Latent Variable,
Given the Observed Scores on Test X



Two-stage latent-model based projection. The link from the distribution of X scores to the distribution of Y scores through the latent variable model is obtained in two steps. First is the sum over examinees of the predictive distributions for θ given their X scores:
$$p(\theta)|\mathbf{X} \approx N^{-1} \sum_i p(\theta|x_i).$$

Two-stage projection

Stage 2: Predictive Distribution for Scores on Test Y,
Given the Predictive Distribution for Latent Variable



The predictive distribution for Y scores is then the sum over examinees of the predictive distributions for Y , convoluted with their predictive distributions for θ , that is,

$$p(\theta) | Y \approx N^{-1} \sum_i \int p(y | \theta) p(\theta | x_i) d\theta.$$

The Matrix

Reporting Scale	Method of Collecting Data				
	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin					
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

Twenty Flavors of Marketbasket Reporting

Each with different implications for what's possible, what's not; what's easy, what's hard; how some things fit with other things.

Reporting Scale	Method of Collecting Data				
	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin					
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

The Easiest Cell in the Matrix


Intuitive Test Theory: What you see is what you get.
But inferences are context bound.

Reporting Scale	Method of Collecting Data				
	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin					
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

A Cell Most People Can Live with

Intuitive Test theory Accepts Equated Scores on Parallel Tests

Reporting Scale	Method of Collecting Data				
	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin					
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					



Flexibility vs Complexity

“Legal Cells” if Projection is Disallowed

Reporting Scale	Method of Collecting Data				
	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin	Yes	Yes	Means Only		
True/Admin	Means Only				
True/Synth	Means Only				
Obs/Synth	Means, if Tau-Equiv				
Theta-hat	Yes	Yes	Maybe, if long or CAT	Maybe, if long or CAT	
Theta-true					

What NAEP is Now

Means and PACs on Latent Variable Scales, “Arbitrary forms.”
This is the second-hardest cell in the matrix to work with.

Reporting Scale	Method of Collecting Data				
	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin					
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

A yellow arrow points from the text "Full Use of Info." to the cell at the intersection of the "Theta-true" row and the "Arbitrary" column. A yellow starburst is positioned to the left of the "Theta-true" row label. The "Arbitrary" column is shaded green.

The Nastiest Cell in the Matrix

We'll see this one again.

Reporting Scale	Method of Collecting Data				
	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin					
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

A yellow starburst is positioned to the left of the 'Obs/Admin' row. A yellow arrow points from the 'Obs/Admin' cell towards the 'Arbitrary' column. A yellow horizontal bar spans across the 'Arbitrary' column, with the text 'Partial Use of Info.' written inside it.

“Administratability” vs Generalizability

Reporting Scale					Arbitrary
Obs/Admin					
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

Convenient to administer, but lacking on content coverage, representation, and generalizability.


“Administratability” vs Generalizability

Reporting Scale	Method of Collecting Data				
	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin					
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

Good generalizability and coverage, but difficult or impossible to administer in standard settings

Dual Reporting Scales

Reporting Scale	Method of Collecting Data				
	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin					
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

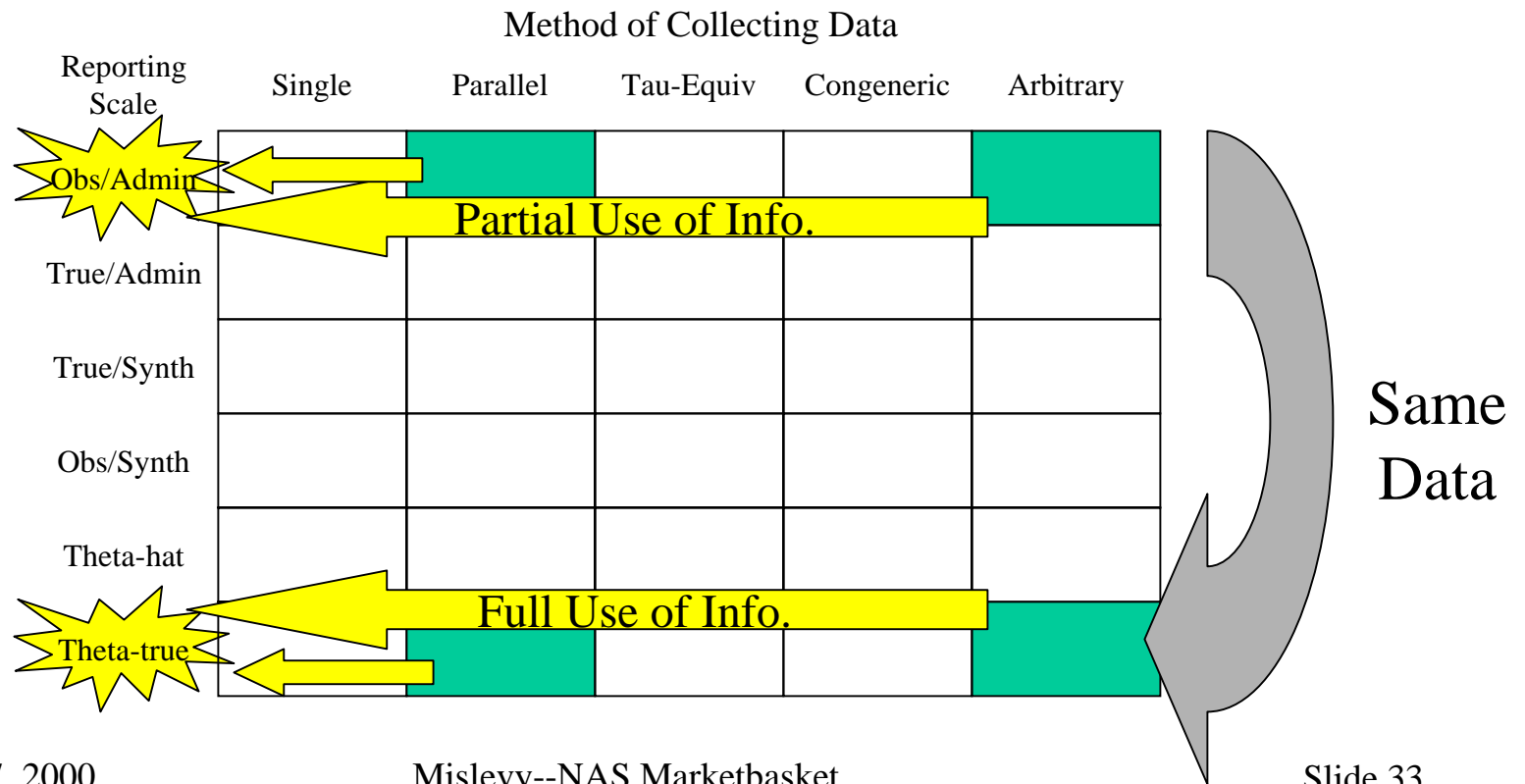


Dual Reporting Scales

Reporting Scale	Method of Collecting Data				
	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin					
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

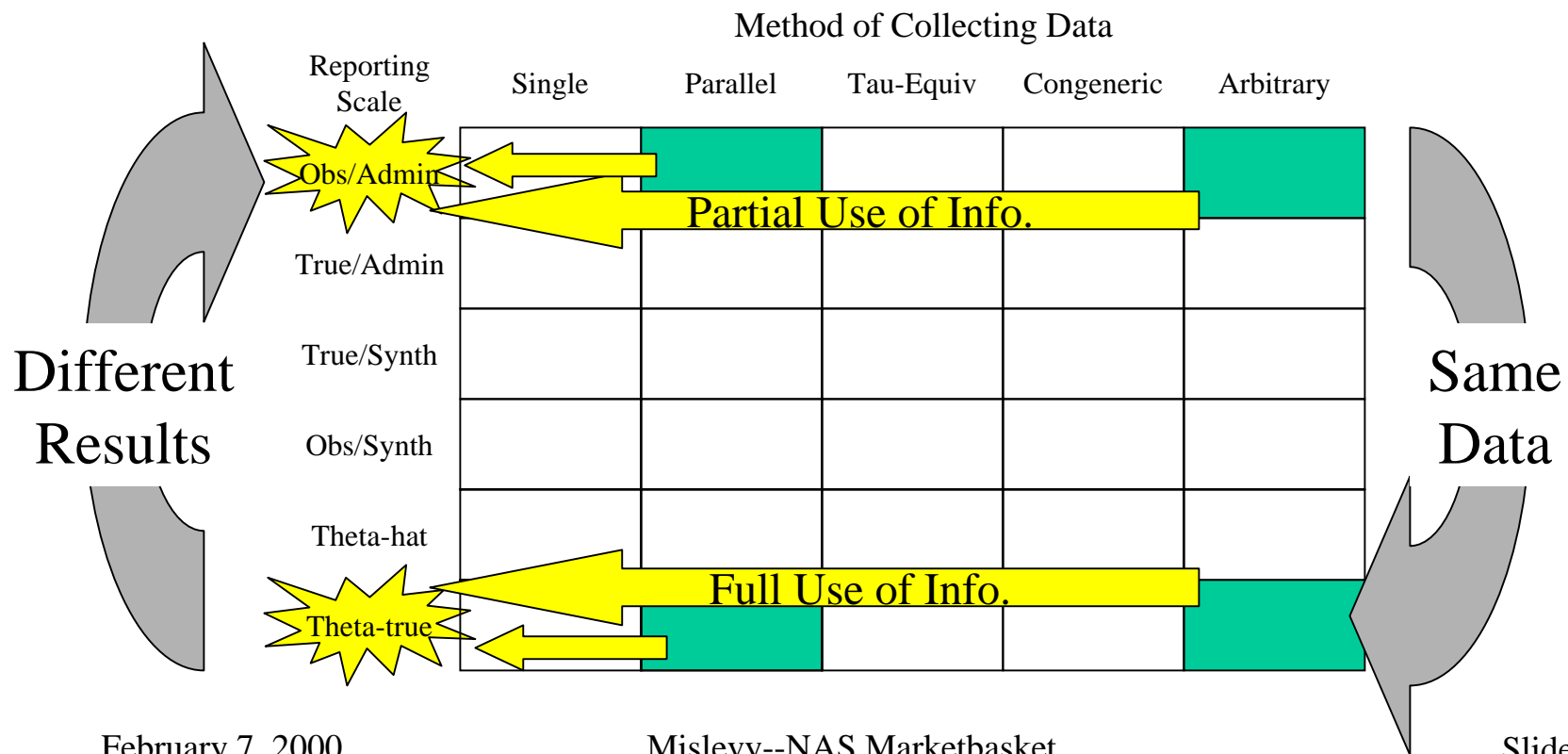
Annotations: A yellow starburst points to the 'Obs/Admin' row. A yellow arrow points from the starburst to the 'Parallel' column. A yellow arrow points from the starburst to the 'Arbitrary' column. A yellow bar with the text 'Partial Use of Info.' spans across the 'Parallel' and 'Arbitrary' columns.

Dual Reporting Scales

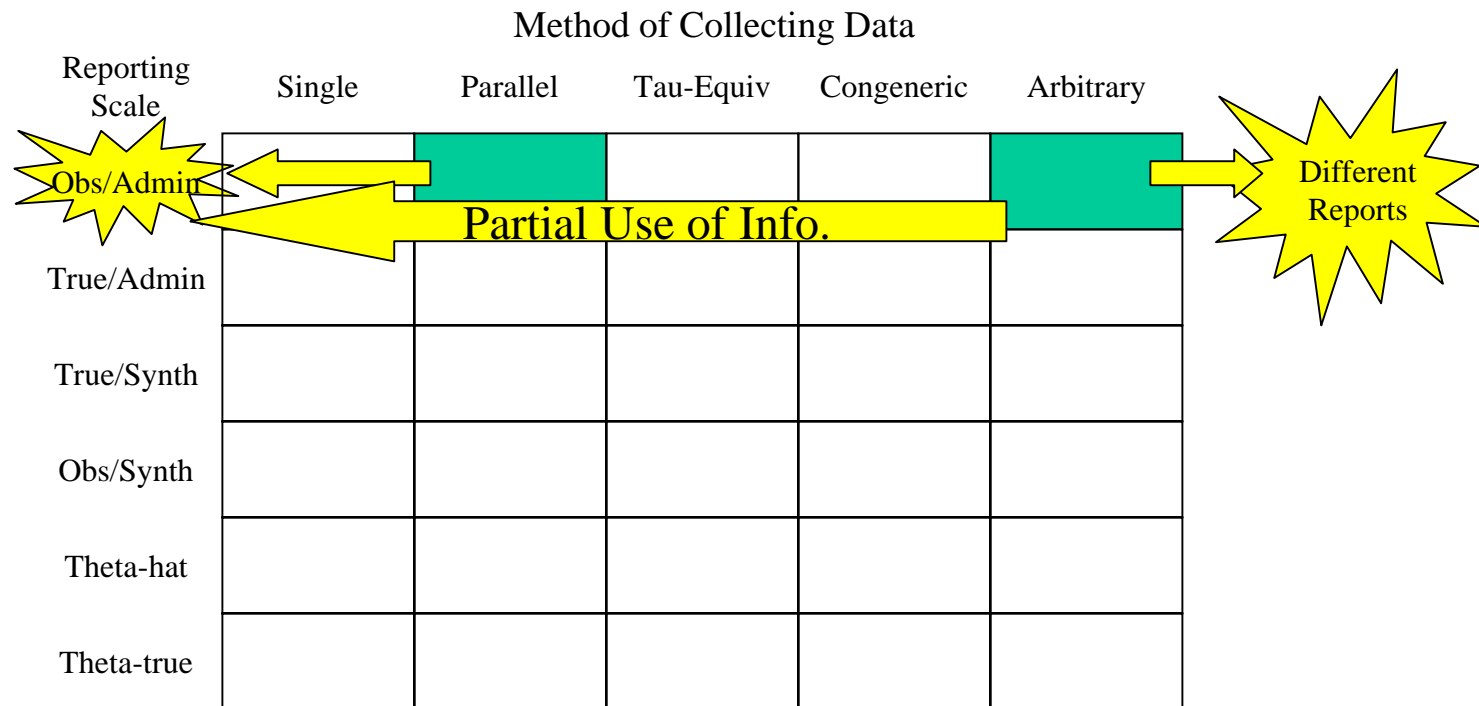


Dual Reporting Scales

Results can disagree for two reasons: True vs Observed, and Different Effective Domains. *PACs especially vulnerable!*



Dual Reporting Strategies



Conclusion: What's Negotiable?

- Methods of data collection are open to negotiation.
- Reporting scales are open to negotiation.
- Targets of inference are open to negotiation.
- Given a combination of data-collection method and reporting scales, the methods of analysis and computing approximations, *among those that account correctly for the evidentiary relationships these choices entail*, are open to negotiation.
- The evidentiary relationships themselves are not.

Implications

- Marketbasket reporting offers a “user-friendly inference” for reporting...
but without serious constraints on data-collection and/or targets of inference, life is not simple behind the scenes.
- Not all desirable combinations of reporting scale, data-collection, targets of inference, and analytic methods are compatible. Sometimes “you can’t get there from here.”
- Make design decisions in light of interactions and tradeoffs, including, importantly, implications for analysis.