Linking Scores from Multiple Instruments

Neil J. Dorans
Center for Statistical theory and Practice
Educational Testing Service

Advances in Health Outcomes Measurement
Bethesda, MD
June 24, 2004

Motivation

The comparability of measurements made in differing circumstances by different methods and investigators is a fundamental pre-condition for all of science.

This statement, made by Dorans and Holland (2000) in the context of equating educational and psychological tests, generalizes to health outcomes measurement.

To the extent that outcome scores of health assessment instruments are to be used interchangeably, the outcome scores need to be equated or made comparable.

Key Questions Addressed

- What is meant by outcome score linking?
- How does equating differ from other types of linking?
- What are common data collection designs used to capture data for outcome scores linking?
- What are some of the standard statistical procedures used to link outcome scores directly?
  What assumptions do they make?
- What role does IRT play in linking outcome scores?
  What assumptions do IRT methods make?

What is Outcome Score Linking?

Equating and linking methods refer to a **collection of techniques** that have been developed by creative individuals to solve the score linking problems that have arisen in a wide variety of practical testing circumstances.

Most of these techniques divide into two categories: **observed** score and **true** score.

Another important distinction is between linear and equipercentile observed score equating methods.

Data collection designs are critical.

When is a Linking an Equating?

In addition to the many techniques for actually doing score linking, there are **five "requirements"** that are often regarded as basic to all of score equating.

Here are **some equivalences I** will be using:

question = item,

instrument = test,

respondent = examinee,

outcome score = score.

The five requirements are:

When is a Linking an Equating?
Equal Construct

1. **The Equal Construct Requirement**: instruments that measure different constructs should not be equated.

   For example, a measure of depression and a measure of anxiety can not be equated.

   Two editions of a depression measure might be equatable.

When is a Linking an Equating?
Equal Reliability

2. **The Equal Reliability Requirement**: instruments that measure the same construct but which differ in reliability should not be equated.

For example, blood pressure readings by a trained professional and those obtained from an index finger machine can not be equated.

Symmetry

3. **The Symmetry Requirement**: the linking function for equating outcome scores of instrument *Y* to those of instrument *X* should be the *inverse* of the linking function for equating the outcome scores of *X* to those of *Y*.

An equation that predicts weight from height will not equal the inverse of the equation that predicts height from weight.

Lord's Equity Requirement

4. **The Equity Requirement**: it ought to be a matter of indifference for a respondent to be tested by either one of two instruments that have been equated.

This requirement has two parts, one practical one not.

The practical part is concerned with average or **expected test performance** and requires equating methods to track and take account of differential test difficulty.

The other part goes beyond means and expected performance. It requires that an examinee ought to have the **same expected distribution** of performance on either one of two equated tests.

It is this aspect of Lord's equity requirement that leads to the pessimistic statement that "equating is either impossible or unnecessary"

When is a Linking an Equating? Population Invariance

5. **Population Invariance Requirement**: the choice of (sub) population used to compute the equating function between the scores of instruments *X* and *Y* should not matter—i.e., the equating function used to link the outcomes of *X* and *Y* should be **population invariant.**

By computing linking functions on subpopulations and comparing them, we can examine and quantify this invariance condition.

Requirement (5) will fail to hold if (1) and (2) do not hold: Linkings between tests that measure different things or are not equally reliable will not be invariant across certain subpopulations of examinees.

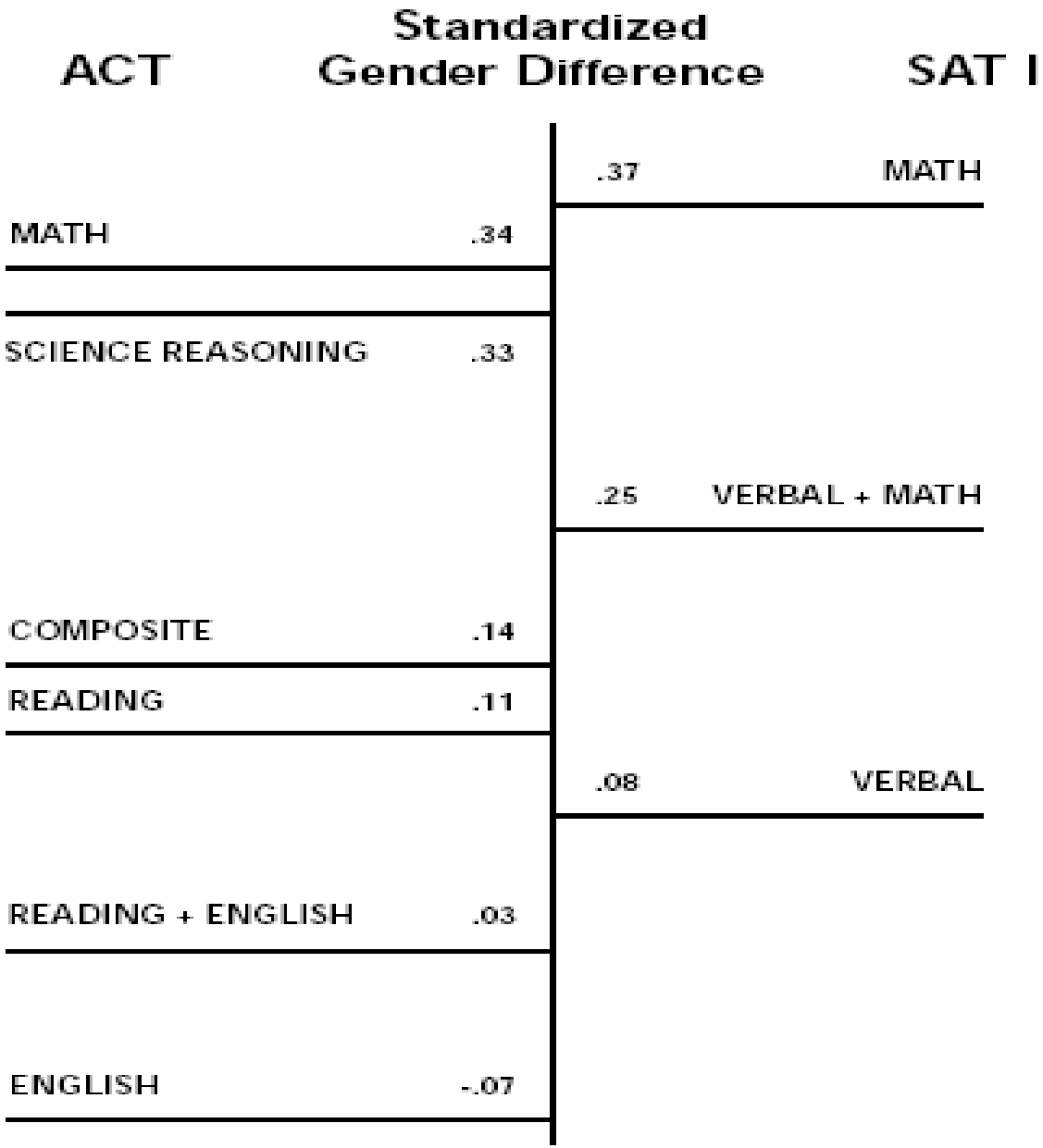
Concordances

Concordance is used to describe links between outcome scores that measure the same or similar constructs but according to **different specifications**.

For example, many colleges and universities accept scores on either the ACT or SAT I. Instead of claiming to equate ACT scores to SAT I scores, a concordance table or concordance function was produced.

This concordance enabled users to better align cut-scores on these two somewhat similar but different tests.

Unlike equatings, concordances are more **sensitive to the population of examinees** whose data are used to estimate the concordance function.


ACT
Standardized Gender Difference
SAT I
MATH .34
SCIENCE REASONING .33
COMPOSITE .14
READING .11
READING + ENGLISH .03
ENGLISH -.07
.37 MATH
.25 VERBAL + MATH
.08 VERBAL

Calibration

- Calibration refers to the process of placing scores on a score scale for tests designed to measure the same construct, but may do so with **unequal reliability** or **unequal difficulty**.
- A content framework is used to ensure that the **construct** being measured is the **same** from one instrument to another.
- A short form of an instrument is less reliable than a longer version and a link between them is an example of a calibration.
- Another example is **vertical linking**, where both instruments may be of similar reliability, but of different difficulty, one being targeted for a different population than the other.

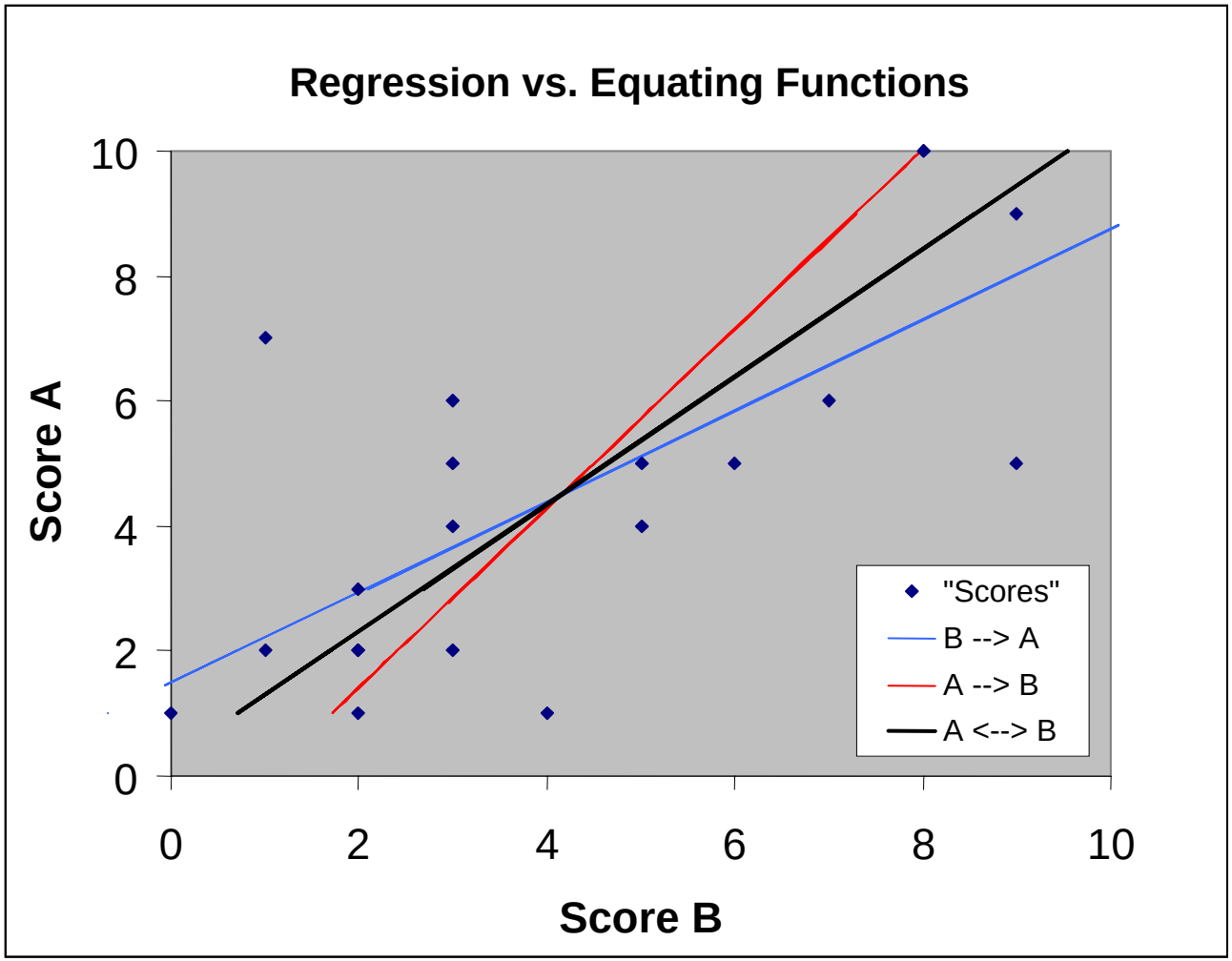
Lack of Interchangeability

- **Equated** scores are **interchangeable**. Concordant and calibrated scores are not interchangeable.
- **Concordances** enable us to **align score scales**, but do not allow us to maintain that the scores can be used interchangeably. Concordances are usually **population dependent.** Height/Weight tables.
- **Calibrations** involve non-equatable measures of the same thing. Often one measure has more score points than the other. Consider the same questions answered yes/no on one test and on 5-point agree/disagree scale on another test.

Prediction of Expected Performance

- A distinction must be made between the methods related to score equating and those of prediction, i.e., regression.
- Both approaches may be used to transform scores on one test into the scale of the scores on another test. However, these transformations have very different uses.
- In **prediction,** the goal is to predict a **expected** *Y*-score for an examinee from some other information about that examinee. In prediction there is an inherent **asymmetry**.
- **Equating** functions do not predict scores on one test from scores on another. Instead, scores that have been equated can be used **interchangeably**. **Symmetry** is essential.

Prediction is Not Equating



Data Collection Designs

The role of data collection is crucial to successful instrument linking.

It is very important to control for differences in distributions of response propensities when assessing differential instrument difficulty.

In test equating or linking, as in most scientific research, this has always been accomplished through the use of special data collection designs.

Types of Data Collection Designs

All types of outcome score linking are based on definitions, data, and assumptions.

Typically data are collected on **complete instruments**, which means that some group of respondents was administered an intact instrument.

Sometimes data on a **complete instrument** is collected in a systematic piecemeal fashion.

Sometimes in a manner that depends on the level of the attribute being assessed - CAT.

Single-Group Design

The single-group design directly controls for differences in response propensities by using the **same respondents for both instruments**.

Special studies, such as those linking the ACT composite to the SAT I V+M score employ this design.

Advantage: groups are very equivalent

Disadvantage: there may be an order effect

Instrument

Population	Sample	***X***	***Y***
P	1	**@**	**@**

The Counterbalanced Design

In order to allow for the **possibility of order effects** in the single group design, the sample is sometimes randomly divided in half and in each sub-sample the two instruments are taken in different orders—*X* first and then *Y*, or *Y* first and then *X*.

This design is rarely employed in practice, but has been used often in special studies that examine relationships between tests built to an old set of specifications and tests built to new specifications.

Instrument

Population	Sample	X_1	X_2	Y_1	Y_2
P	1	**@**			**@**
P	2		**@**	**@**	

The Equivalent Groups Design

In the equivalent groups design, two **equivalent samples** are taken from a common population **P**; one is administered instrument *X* and the other instrument *Y*.

Obtaining **large representative** equivalents groups is the key to success with this design.

ACT employs this design to place new ACT forms on the 1-36 scale.

Instrument

Population	Sample	***X***	***Y***
P	1	**@**	
P	2		**@**

Strong Data

These three single population designs yield strong data because differences in distributions of responses are eliminated directly by constructing equivalent groups.

In essence, alternate instruments are randomly assigned to equivalent groups of respondents, as in randomized experiments.

Other designs produce weaker data that require more assumptions to produce linking relationships.

Anchor Instrument Designs (NEAT)

The anchor instrument designs improved upon the flexibility of the single population designs, by allowing the two samples, one from population **P** that is given *X* and one from population **Q** that is given *Y*, to be different or "non-equivalent."

However, the two samples must only be different in ways that can be quantified using an anchor instrument, *A*, which is administered to both **P** and **Q**.

Population	Sample	***X***	***A***	***Y***
P	1	**@**	**@**	
Q	2		**@**	**@**

Anchor Instrument Designs (NEAT)

Because *X* is never observed for examinees in **Q**, and *Y* is never observed for examinees in **P**, **some type of assumption is required to "fill in" these missing data.**

For this reason, there are **more methods of equating instruments for anchor instrument designs** than there are for the others. All of these methods correspond to assumptions made about the missing data.

This design is used to place new SAT I forms on the 200-800 scale.

Incomplete Instrument Designs

The anchor instrument design collects data for linking instruments that are administered to different groups of respondents.

There exist designs for linking instruments that have never been administered to the same set of respondents.

These designs range from the highly structured data collections for section pre-equating to designs used for computer adaptive testing. These less structured designs produced weaker data that require stronger assumptions in order to produce links.

Instrument

Group	U_1	C_1	U_2	…	U_K
G_1	@	@			
G_2			@		
G_3					
G_4		@			
:				@	@
G_J			@		

Incomplete Instrument Designs

The number of questions associated with a unique set of questions (*U*) or a linking set (*C*) can vary from a handful to a large number. The instruments to be linked may be composed of questions from different *U*s and *C*s.

The groups (**G**) may come from the same population or many different populations.

These less structured designs produce weaker data that require stronger assumptions in order to produce links. IRT models come in handy here.

Observed Score Equating Definitions

Linear linking definition:

X-score *x* and *Y*-score *y* scores corresponding to the same number of standard deviations above or below the mean in population **T** are equivalent

Equipercentile definition:

X-score *x* and *Y*-score *y* are *linked* in **T** if $F_{\mathbf{T}}(x) = G_{\mathbf{T}}(y)$. When these two cdf's are continuous and strictly increasing, then this equation can always be satisfied.

Equivalent Groups Linear (MSD)

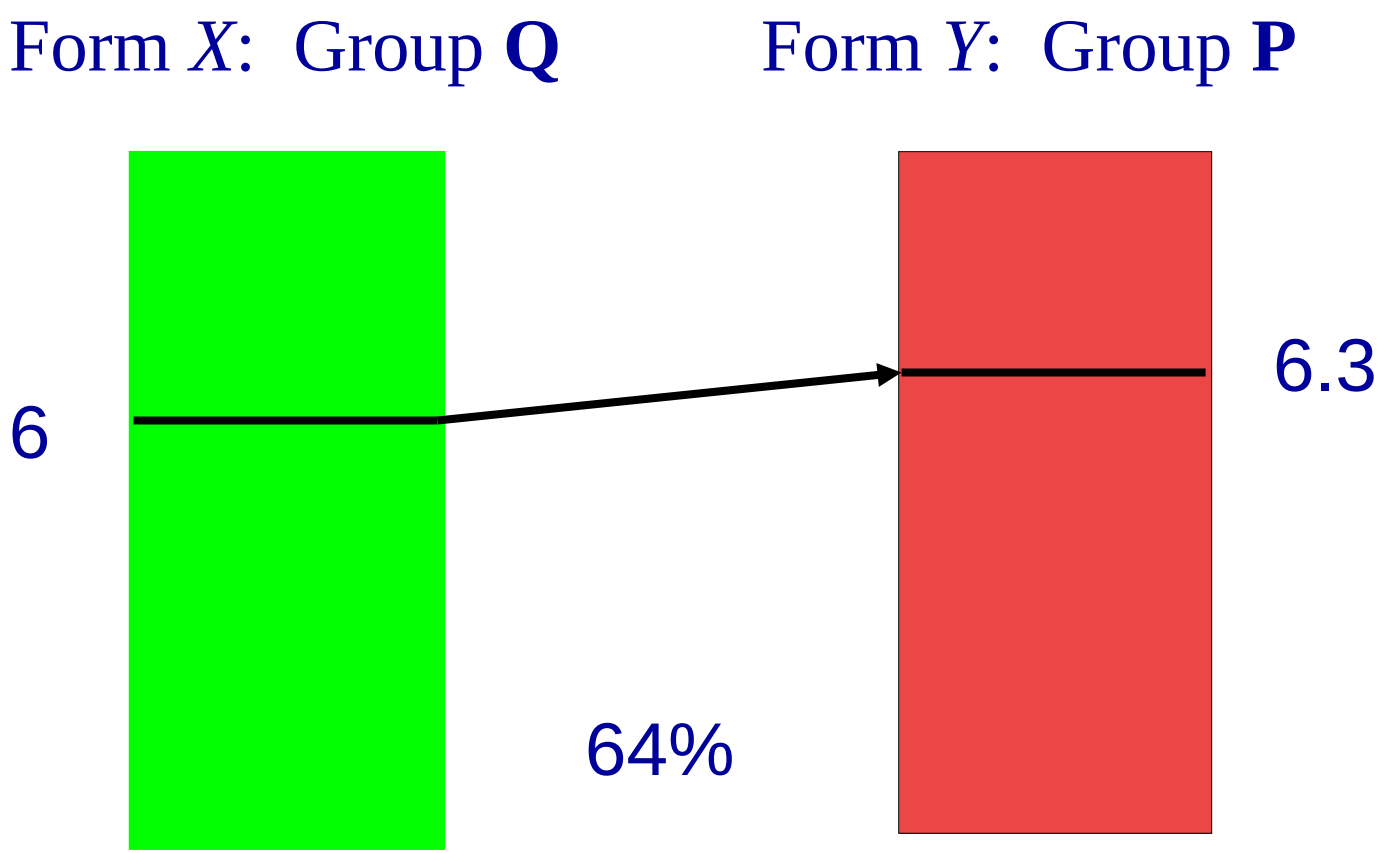
- Set mean and standard deviation of the new form X equal to the mean and standard deviation of the old form Y.

- Place the raw scores X on the raw score scale for Y using the linear equating function:

$$X_Y = (S_Y / S_X) * (X - M_X) + M_Y$$

  - where M and S represent mean and standard deviation

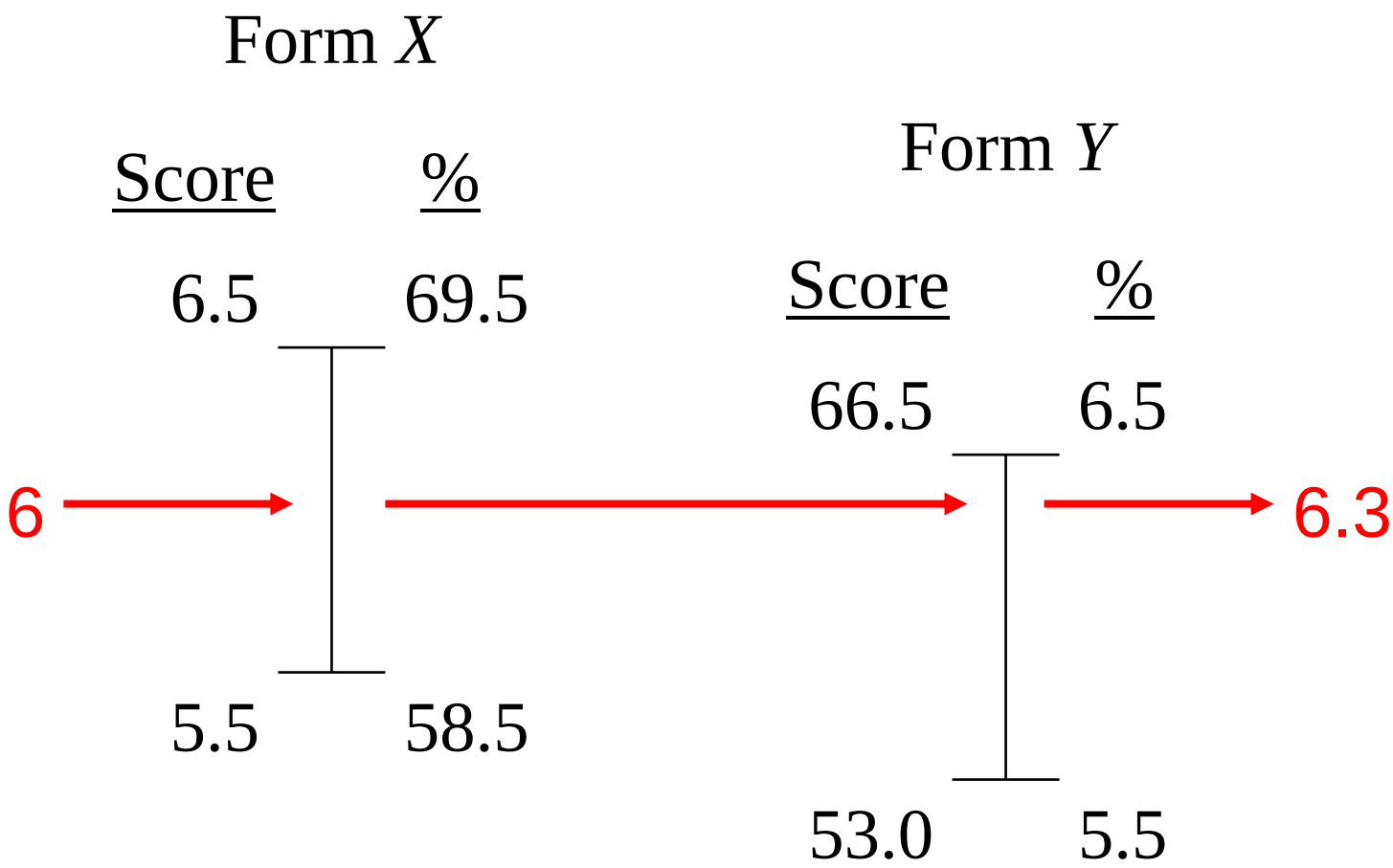
Equipercentile Equating



A Closer Look

Raw Score	Cum % below Form *Y*	Cum % below Form *X*
0	2.0	4.5
1	8.0	11.0
2	16.5	20.0
3	26.5	32.5
4	41.0	42.0
5	53.0	58.5
6	66.5	69.5
7	77.0	84.5
8	89.5	93.0
9	96.0	96.0
10	100.0	100.0

A Closer Look



Equipercentile Results: Form *X* on Form *Y* Scale

Raw Score on *X*	Equated Score on *Y*
0	0.9
1	1.5
2	2.4
3	3.5
4	4.2
5	5.3
6	6.3
7	7.5
8	8.4
9	9.3
10	10.0

Equipercentile Results: Form *X* on Score Scale

Raw Score on *X*	Scaled Score
0	109.1
1	114.6
2	123.8
3	134.8
4	142.4
5	152.7
6	163.2
7	175.0
8	184.4
9	192.7
10	200.0

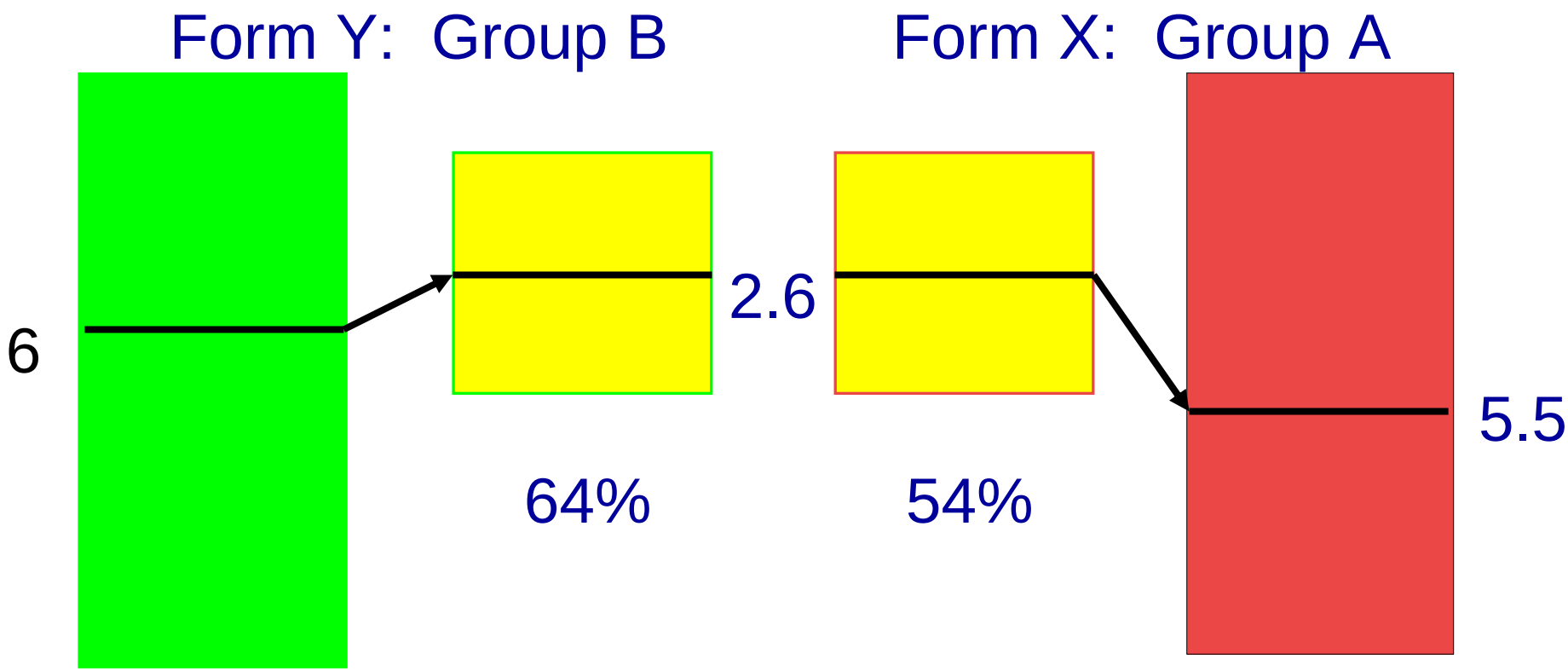
NEAT Designs – Non-Linear Methods

Post-stratification: Assume that the conditional distribution of the instrument given the anchor is the same in the full population **T** as it is in the sample from the subpopulation (**P** or **Q**) where it is seen. (**Invariant total/anchor regressions**).

Chained: Assume the that linking between the anchor and the total test is the same in sample as in the full population. (**Invariant calibrations**).

IRT: Assume that the **item response function** for each item is **invariant** across all subpopulations.

Chained Equipercentile Equating



Linear Equating Methods – NEAT Designs

Chained linear assumes that the **mean/sigma linking** relationship between A and X is population **invariant**, as is the the mean/sigma linking relationship between Y and A.

Tucker linear assumes that the **best linear predictor** of Y from A is population **invariant**, as is the best linear predictor of X from A**.**

Levine equally reliable linear equating model assumes that the **true scores** on X, Y and A are **perfectly related**, and that X and Y are **equally reliable**.

Note that when the correlation between the anchor and the total tests equals one, all three linear methods converge.

IRT's Role in Linking Instruments

Item response theory can be used to **link data collected with any design** ranging from the strongest single group design to the weakest design in which only handfuls of items are administered to the same group of people.

Its **flexibility** is a direct consequence of its **strong assumptions.**

IRT produces **indirect outcome score linking** as opposed to direct linking associated with some observed outcome score linking methods.

IRT Scale Linking

In IRT linking occurs at the level of the metric for ability and item parameter estimates.

For NEAT designs or incomplete instrument designs, the metrics are linked either through

- Joint estimation of item and people parameters
- Estimation of items and people parameters in separate estimation runs that share common anchor questions that can be used to link the metrics via linear equations.

von Davier & von Davier (2004) describe a unified approach to IRT scale linking and scale transformations (ETS RR –04-09)

IRT Equating

True-score (Lord, 1980; p. 199-201):

True scores (sum of IRFs) on instruments that correspond to the same latent ability are equated. Regressions of observed scores onto latent ability can serve as equating functions for true scores.

Observed-score (Lord, 1980; p. 202-203):

Use estimated latent ability distribution in **T**, estimated IRFs, and the generalized binomial theorem to estimate joint distribution of instrument scores, and equipercentile equate.

Weak Data and Complex Models

Weak data requires complicated models with their many assumptions.

Anchor instrument equating models make assumptions about instrument outcome scores and anchor instrument outcome scores.

IRT models make strong item level assumptions that make it potentially useful in a variety of settings.

Invariance assumptions are made at the level of regressions, linkings or item parameters.

Strong Assumptions of IRT

IRT makes **strong assumptions** at the **level of questions** that establishes linkages at the question level.

From these question level linkages, **indirect linkages** among instruments outcome scores can be constructed.

IRT models provide parameterization of the item space and person space that produce **item parameter invariance** across subpopulations of examinees **if the model fits**.

Many models presume that there is only a single person parameter is needed and that this person parameter combines with a set of item parameters the describe examinee performance at the item level.

Check Assumptions via Equity Checks

Some function, be it the IRF of IRT or the equipercentile equating function or the regression of total score on anchor score, is assumed to be invariant across subpopulations.

Males and Females can be used to test for equatability.

Estimate functions in both groups and check for invariance across populations.

The current issue of *Journal of Educational Measurement* examines this population invarince issue with Advanced Placement data.

Summary

Equating was defined and contrasted with other forms of outcome score linking. Equating is only approximated in practice because its strong requirements are hard to meet, e.g., population invariance of linking functions.

Different data collection designs were described along with some methods used with these designs.

Relatively assumption-free simple models can be applied to the strong data collected within a single population (single- or equivalent- group designs).