SCANTRON®

# How Differential Item Functioning Analysis (DIF) Can Increase the Fairness and Accuracy of Your Assessments

Nikki Eatchel, SVP of Assessment
David Grinham, SVP International Assessment Solutions

# Scantron At A Glance

### The Iconic Brand in Assessment

SCANTRON®

- 12 billion assessments since 2000 — 100+ million digital assessments
- 40+ year track record of accuracy and reliability

### Expansive Scale, Reach, and Reputation

- Provider to Fortune 500 companies, government institutions, & 48 Ministries of Education
- Providing products and services in 65 countries

### Assessment Development

- Support for both paper-based and online testing
- Psychometric expertise in IRT, form-based and adaptive testing, and the full assessment development cycle

### Reporting and Analytics

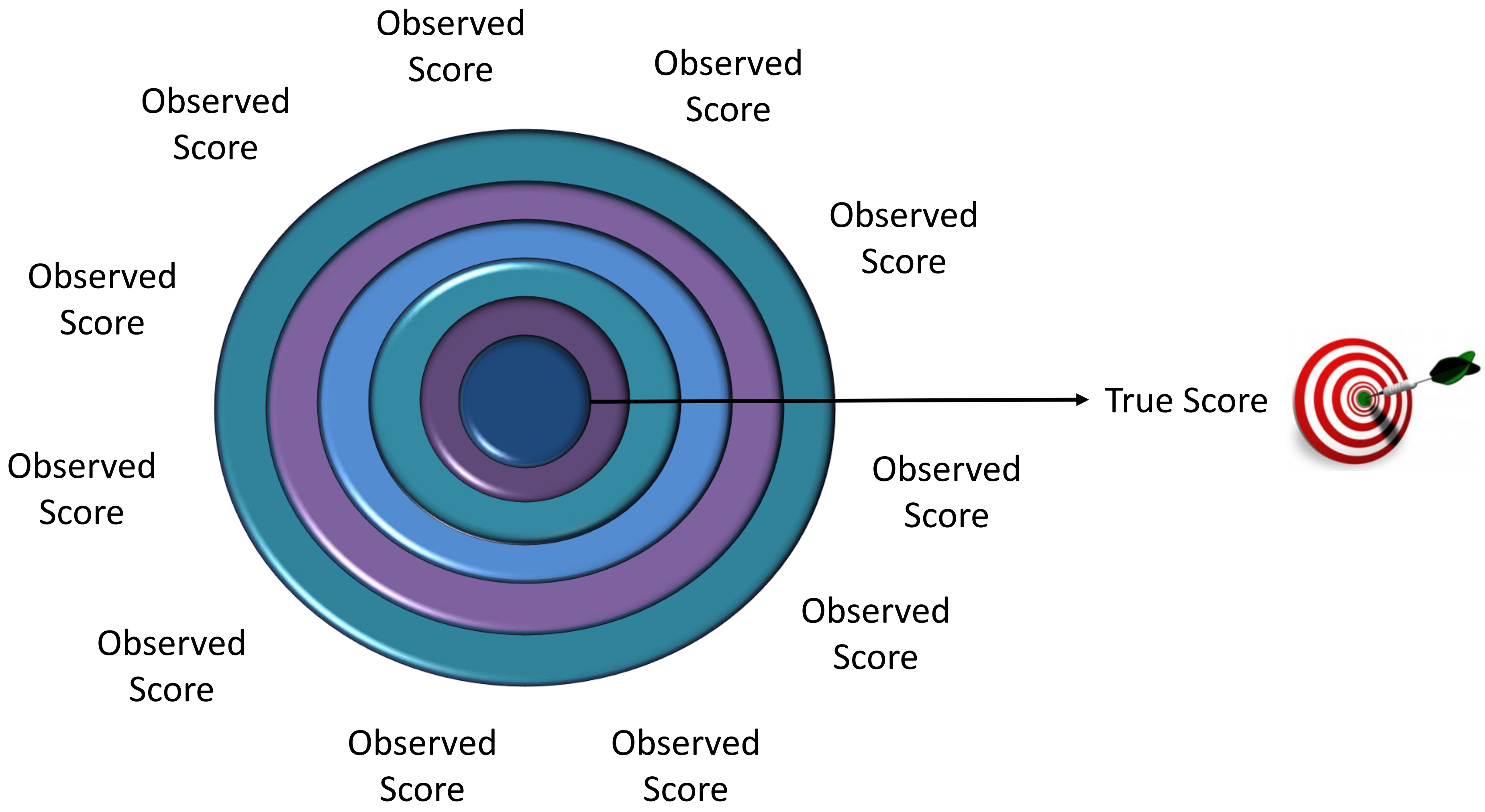
- Valid and reliable reporting for large-scale programs
- Innovative analytics tools to evaluate and inform program decisions

# The Imperative for Fairness and Validity in Assessment and Selection

# The Imperative for Fairness & Validity

Bias within selection (even when unintentional) impacts the fairness and validity of the process and can result in a truncated talent pool and a lack of equal opportunity for candidates.

# It's About Reducing Error



Observed Score

Observed Score

Observed Score

Observed Score

Observed Score

True Score

Observed Score

Observed Score

Observed Score

Observed Score

Observed Score

Observed Score

# How Do You Accomplish Error Reduction?

Professionals in the assessment industry are continually focused on one main issue – the reduction of error.

Scores are never without error, but error can be reduced for all candidates through a variety of assessment development and assessment evaluation techniques, including:

- Universal Design
  Sensitivity/Bias Reviews
- Enhanced Translation Procedures
- Differential Item Analysis

Development and translation activities reduce the error in scores

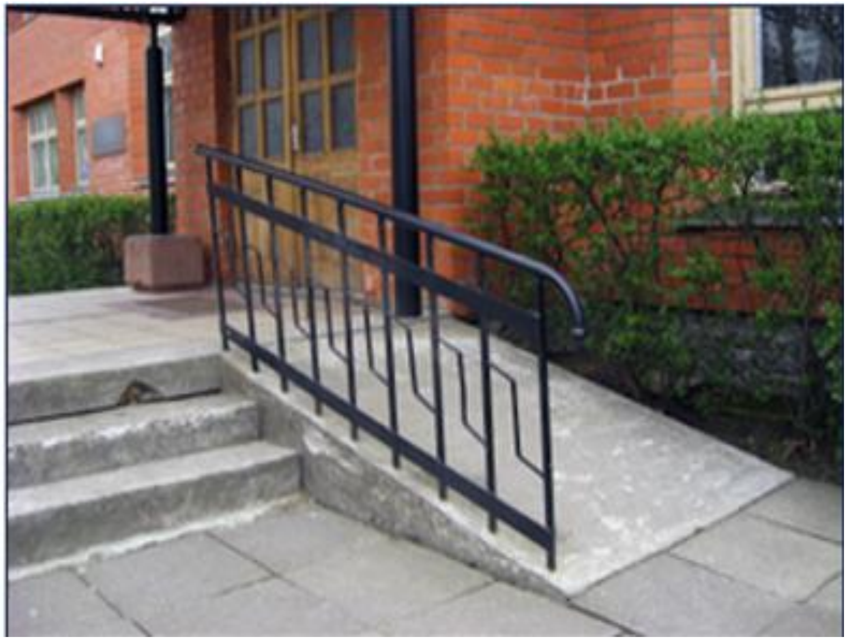
# Universal Design

# What is Universal Design?

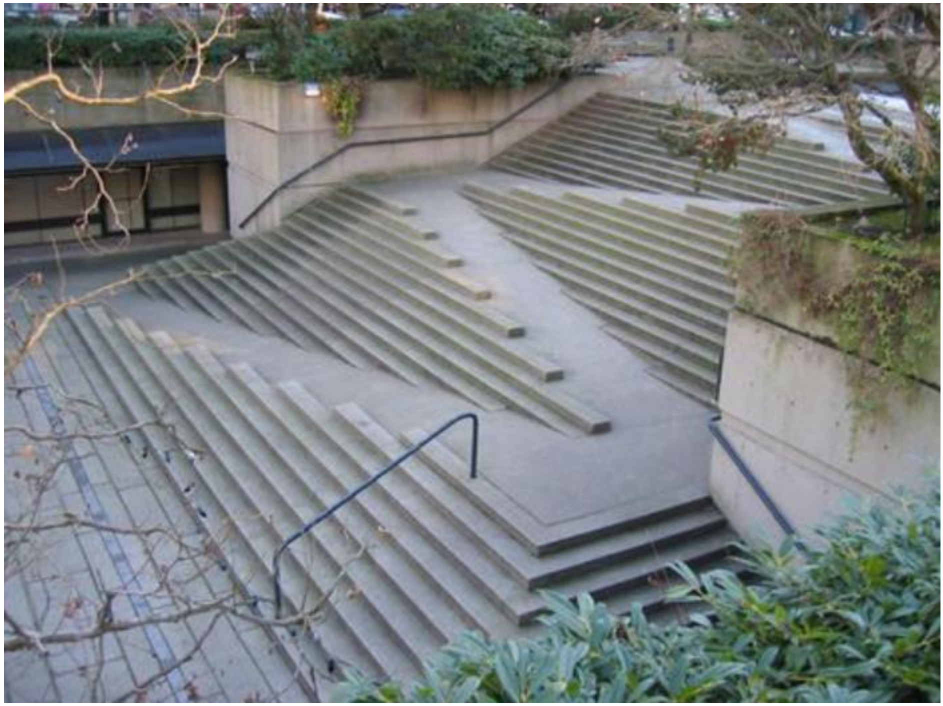
Universal design is a concept that began in the field of architecture. It has expanded into many other realms, including education.

Defined by the Center for Universal Design as “the design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design.” (1997)

# What is Universal Design?



***Before*** **universal design**



***After*** **universal design**

# 7 Principles of Universal Design in Assessment

1. Inclusive assessment population
2. Precisely defined constructs
3. Non-biased items
4. Amenable to accommodations
5. Simple and clear instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

# Sensitivity/Bias Reviews

# Why conduct sensitivity and bias reviews?

Sensitivity and bias reviews are an essential step within the assessment development process. Item content (text, graphics, passages, scenarios, videos, etc.) should be evaluated to determine its familiarity and appropriateness for different subgroups:

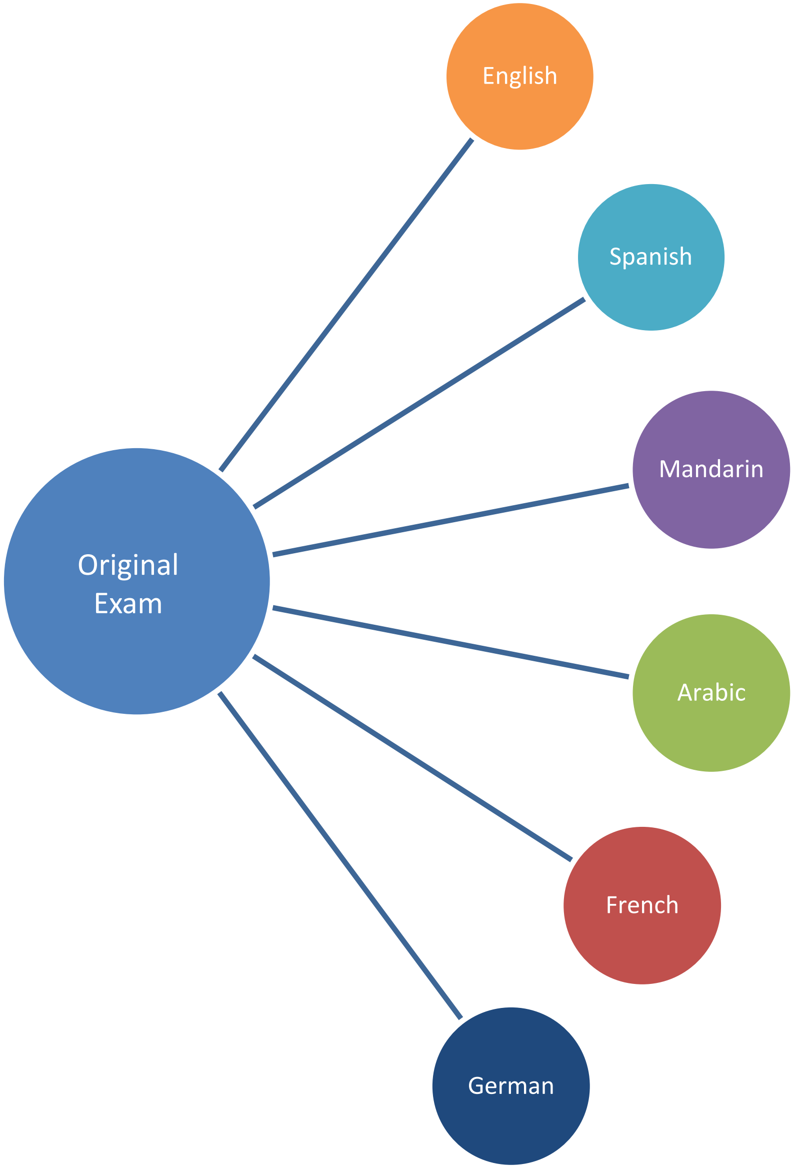
- Cultural
- Geographic
- Gender
- Ethnicity
- Socioeconomic status

# Translation and Localization

# Increase of Available Assessment Languages

To address one of the extraneous variables impacting candidate scores (native language), many assessments are translated/adapted into additional languages.

While this is a positive step in candidate fairness and equal access, it comes with additional concerns and considerations.



# Translation/Localization Considerations

As items are translated, there are several important considerations that could potentially impact the validity of your exam scores:

- Is the same construct measured across the different language versions of the test?
- Is the same level of knowledge and skill required to answer the item correctly consistent across languages?
- Are there items that perform differently across the various language versions of the test?

# Differential Item Analysis (DIF)

# Differential Item Analysis - Definition

A test item is flagged within a DIF analysis when examinees with equal ability, but from different groups, have an unequal probability of performing successfully on the item.

The analysis uses a “reference group” versus a “focal group.”

- Males, females
- White, blacks
- English, Spanish

# Item Impact, DIF, and Item Bias

**Item Impact**: When one group of candidates performs significantly differently (higher or lower) on an item than another group.

**Differential Item Functioning**: Statistically evaluates whether item impact is the result of overall group differences in proficiency.

**Item Bias**: A qualitative evaluation of whether group differences in performance are based on variables irrelevant to the construct the test is intended to measure.

*Appraising item equivalence across multipole languages and cultures*, Sireci & Allalouf, National Institute for Testing and Evaluation, 2003

# DIF versus Bias: Practical Examples

# Methods of Analyses

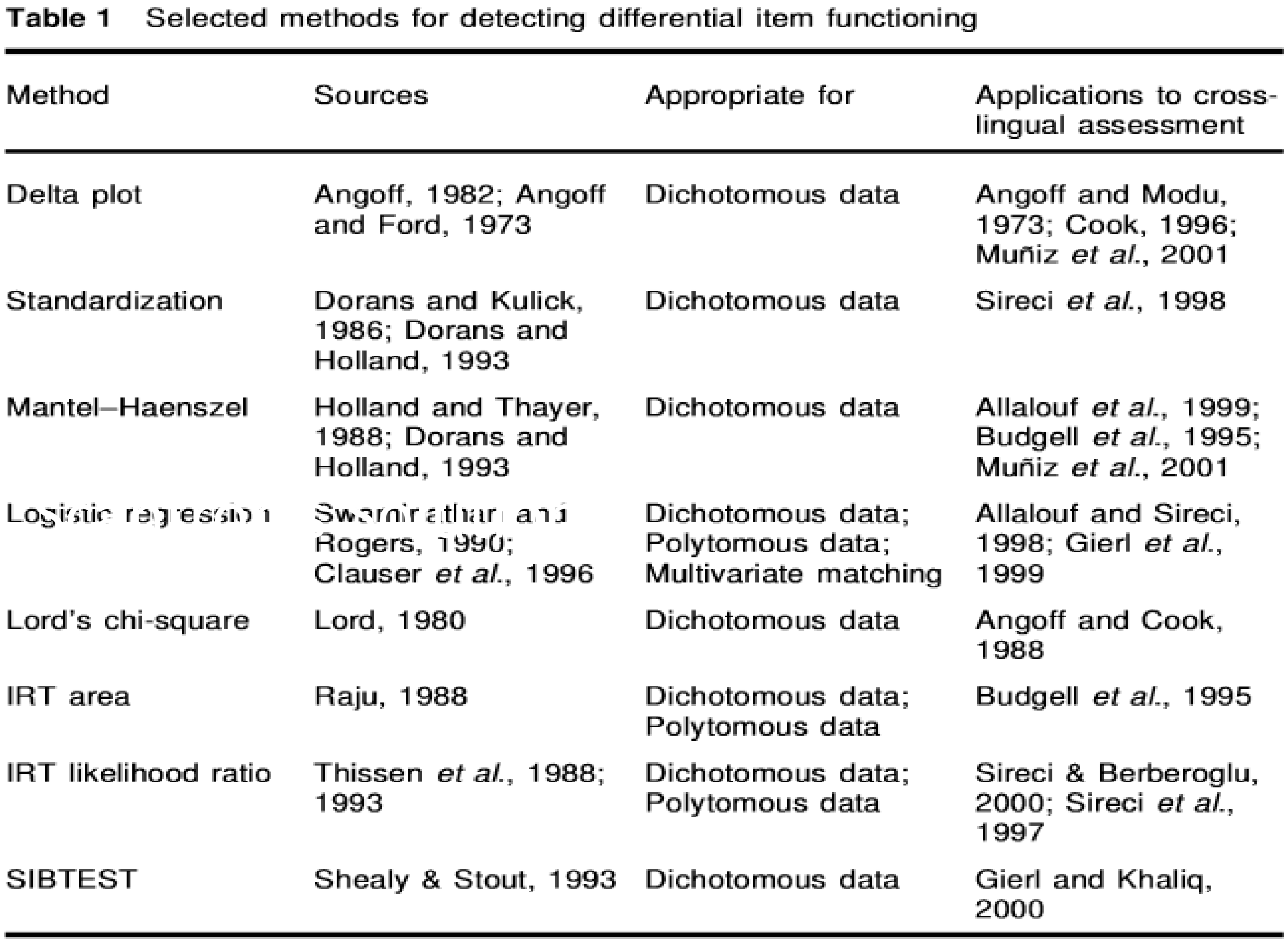


**Table 1** Selected methods for detecting differential item functioning

| Method | Sources | Appropriate for | Applications to cross-lingual assessment |
|---|---|---|---|
| Delta plot | Angoff, 1982; Angoff and Ford, 1973 | Dichotomous data | Angoff and Modu, 1973; Cook, 1996; Muñiz *et al.*, 2001 |
| Standardization | Dorans and Kulick, 1986; Dorans and Holland, 1993 | Dichotomous data | Sireci *et al.*, 1998 |
| Mantel–Haenszel | Holland and Thayer, 1988; Dorans and Holland, 1993 | Dichotomous data | Allalouf *et al.*, 1999; Budgell *et al.*, 1995; Muñiz *et al.*, 2001 |
| Logistic regression | Swaminathan and Rogers, 1990; Clauser *et al.*, 1996 | Dichotomous data; Polytomous data; Multivariate matching | Allalouf and Sireci, 1998; Gierl *et al.*, 1999 |
| Lord's chi-square | Lord, 1980 | Dichotomous data | Angoff and Cook, 1988 |
| IRT area | Raju, 1988 | Dichotomous data; Polytomous data | Budgell *et al.*, 1995 |
| IRT likelihood ratio | Thissen *et al.*, 1988; 1993 | Dichotomous data; Polytomous data | Sireci & Berberoglu, 2000; Sireci *et al.*, 1997 |
| SIBTEST | Shealy & Stout, 1993 | Dichotomous data | Gierl and Khaliq, 2000 |

*Appraising item equivalence across multiple languages and cultures*, Sireci & Allalouf, National Institute for Testing and Evaluation, 2003

# 2 Methods of DIF Analyses: Delta Plot

The Delta Plot Method for evaluating cross-cultural DIF is relatively easy to do and is easy to interpret.

Caution:

Misses items if the discriminating power at the high and low end of the scale.

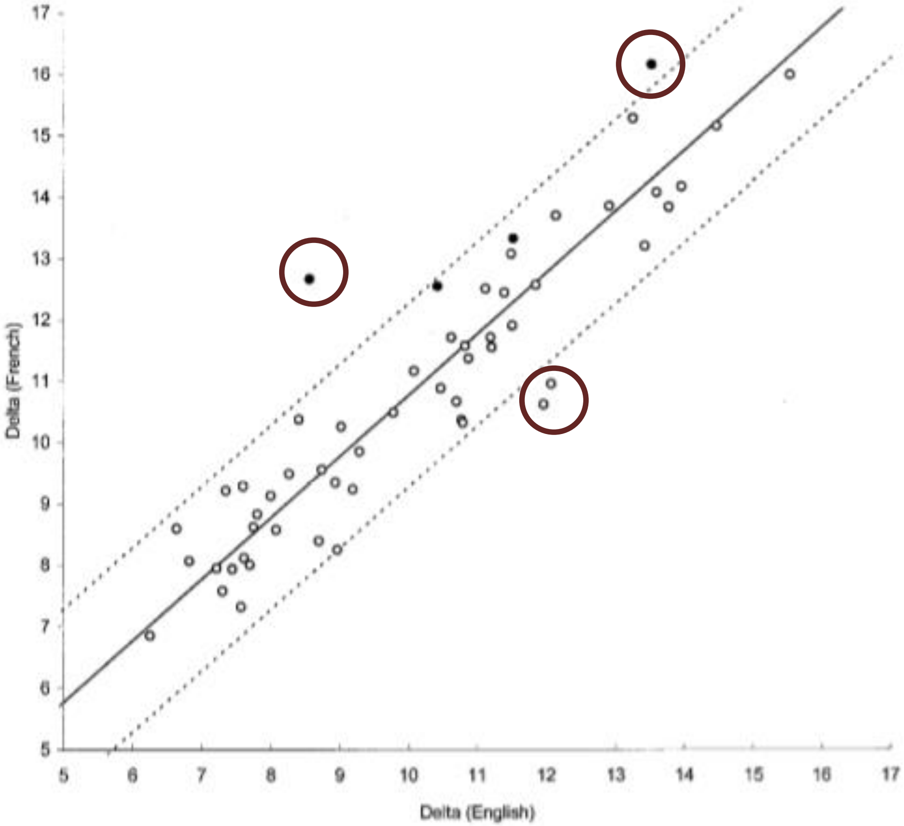


**Figure 1** Plot of English ($n = 2000$) and French ($n = 1333$) group delta values (with mean difference .77 adusted)
*Source*: Muñiz *et al.*, 2001; *Note*: DIF items are represented by black dots.

*Appraising item equivalence across multiple languages and cultures*, Sireci & Allalouf, National Institute for Testing and Evaluation, 2003

# 2 Methods of DIF Analyses: M-H Method

The Mantel-Haenszel method for identifying DIF matches candidates from two different groups on the proficiency of interest and then compares the likelihood of success for the two groups across the score scale, using three sets of information:

- Examinee group (e.g., two different language groups)
- Matching variable (scores upon which examinees in a different groups are matched)
- Individual item response (correct or incorrect)

*Appraising item equivalence across multiple languages and cultures*, Sireci & Allalouf, National Institute for Testing and Evaluation, 2003

# Example Studies

# Assessment Study: Language Groups

Evaluation of multiple language groups: Hebrew and Russian

- Statistical identification of 42 DIF items
- Determination of which group DIF favored
- Review of DIF and non-DIF items by translator group for any potential translation issues

*Appraising item equivalence across multiple languages and cultures*, Sireci & Allalouf, National Institute for Testing and Evaluation, 2003

# Assessment Study: Language Groups

Items with DIF had the following translation issues:

- 38.1% of the DIF items had changes in difficulty of words or sentences due to translation
- 19% of the DIF items had changes in content, effectively creating different items.
- 14% of items had differences in the relevance of the content of the item to each culture
- 12% of the DIF items had changes in the format of the item.

*Appraising item equivalence across multiple languages and cultures*, Sireci & Allalouf, National Institute for Testing and Evaluation, 2003

# Job Analysis Study: Language Groups

Evaluation of cross-language DIF on the Job Content Questionnaire (JCQ)

- 27 items evaluated in six languages (French, Dutch, Belgian-French, Belgian-Dutch, Italian, and Swedish)
- On average, 36% to 39% of the total tested items appeared to be cross-language DIF items.
- Panel review indicated that half of the DIF items may be associated with translation differences.
  - Missing/added words
  - Changes in item difficulty

*Cross-Language Differential Item Functioning of the Job Content Questionnaire Among European Countries: The JACE Study*, Choi et. al, 2009

# Program Considerations

# Major Steps for a DIF Study

- Identify Reference and Focal Groups of interest
- Design the DIF study to have the largest samples sizes possible
- Choose DIF statistics that are appropriate for the data
- Conduct the analyses (one or more as desired)
- Review DIF findings with panel of SMEs
- Determine action per item

# But before you start . . .

- Build DIF into your budget
  - Psychometricians
  - SME Panel (for DIF review)

- Consider your timing
  - Pretest/field test evaluation
  - Operational evaluation
  - Development cycles
  - Retirement cycles

# But before you start . . .

- Consider implications
  - DIF results need to be addressed in a timely manner
  - Development cycles (ongoing maintenance) may need to increase to cover item retirement due to DIF
  - DIF findings should be use to improve both the development and translation process

# Questions?

# References

*Adapting Credentialing Examinations for International Uses*, S. Sireci et. al, 1998

*Appraising item equivalence across multiple languages and cultures*, Sireci & Allalouf, National Institute for Testing and Evaluation, 2003

*Cross-Language Differential Item Functioning of the Job Content Questionnaire Among European Countries: The JACE Study*, Choi et. al, 2009

*Ensuring Validity of NCLEX® With Differential Item Functioning Analysis*, Woo & Dragan, Journal of Nursing Regulation. 2012

*ITC Guidelines for Translating and Adapting Tests,* ITC, 2005

*Standards for Educational and Psychological Testing*, AERA, APA, NCME. 2014.