Statistical and Qualitative Approaches for Facilitating Comparability in Cross-Lingual Assessments

Stephen G. Sireci
Center for Educational Assessment
University of Massachusetts, USA

Presentation delivered at the OECD Seminar on Translating and Adapting Instruments in Large-Scale Assessments
June 8, 2018, Paris

Given that I do not have time to talk about all of the things I would like to discuss today, I will start by giving my take-home messages.

Take-home messages

- **Cross-lingual assessment is hard!**
- **We can never achieve full equivalence of different-language versions of assessments.**
- **Evaluation of the validity of adaptations depends on testing purpose.**
- **For most purposes, we can achieve sufficient validity evidence.**
- **MDS is an under-used method for evaluating equivalence**
- **Many interpretations need to be qualified (“interpreted cautiously”)**

And one new (radical?) idea

- **Develop indices of adaptation comparability**
  - **Describe level of confidence in making "comparative" inferences.**

Multiple-language versions of a test

- **Seen as one way to promote *FAIRNESS* by allowing examinees to access and interact with the test in their native language.**

The *Standards for Educational & Psychological Testing*

A test that is fair within the meaning of the *Standards*

a) reflects the same construct(s) for all test takers

b) scores from it have the same meaning for all individuals in the intended population

c) a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct.

Interest in "multilingualism" in the USA is very different from 100 years ago:

"If English was good enough for Jesus, it's good enough for the school children of Texas."

Texas Governor James "Pa" Ferguson (1917) after vetoing a bill to finance the teaching of foreign languages in classrooms.

How do we assess people who communicate using different languages?

- **The most common procedure is to translate (adapt) an existing test into other languages.**

In this talk, I will describe

1. **validity issues**
2. **“quality control” procedures**
3. **research designs**
4. **statistical methods**

for developing and evaluating tests designed for cross-lingual assessment

And briefly mention

5. **Standards and Guidelines relevant to cross-lingual assessment**
   - **AERA et al. (2014) *Standards***
   - **ITC *Guidelines on Translating and Adapting Tests* (2017)**
   - **Briefly (Hambleton, 2018—today!)**

STANDARDS
for Educational and Psychological Testing

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION
AMERICAN PSYCHOLOGICAL ASSOCIATION
NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

of the test, or that the
roduces scores that
ble/precise and valid...

ITC "Test Adaptation" Guidelines (2017)

- **Test developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the language versions of the test, and (2) identify problematic components or aspects of the test that may be inadequate in one or more of the tested populations.**

SUMMARY: What do the professional standards say?

- **Adapted tests cannot be assumed to be equivalent.**
- **Research is necessary to demonstrate test and score comparability.**
- **Statistical methods can help evaluate item/test comparability.**

Test Adaptation and Research Methodology

Good research designs are needed for

- **The translation/adaptation process (qualitative)**
- **Evaluating the similarity (comparability) of the scores from different language tests (quantitative)**
- **Establishing formal relationships among the different tests (quantitative)**

Adaptation and Validity

- **All research designs are designed to provide evidence of validity**
  - **of score interpretations**
  - **for the use of the test scores**

But what is *validity*?

Standards' Definition

- "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests"

AERA, APA, & NCME (2014, p. 11)

Therefore, validity issues in cross-lingual assessment must consider the *purpose* of the assessment

- **Validation involves gathering data to support (defend) the use of a test for a particular purpose.**
- **Different research designs are needed to support different *uses* of test scores (i.e., different interpretations of scores)**

Adapting a test to use in another language/culture

- **If test score interpretations are to be made WITHIN a language group, no evidence of comparability across languages is needed.**
  - **Evidence that scores are appropriate for whatever they are used for is (obviously) still needed.**
  - **E.g., Scores on the Prueba de Aptitud Academica used only for selection into graduate school in Puerto Rico**

However, if scores are to be compared ACROSS language groups:

- **Need to validate *comparative* inferences**
  - **E.g, TIMSS, PISA**
- **How?**
  - **Demonstrate construct equivalence**
  - **Rule out method bias**
  - **Rule out item bias**

(van der Vijver)

Are scores from different language versions of an exam supposed to be "comparable?"

- **If comparable means "equated," or on the same scale, this is probably impossible.**
  - **In its strictest sense, equating implies examinees would get the same score no matter which "form" of the exam they took.**

However, comparability does not need to mean "equivalent" or equated.

Can test scores from different language versions of an exam be considered “comparable?”

- **Through (a) careful adaptation procedures, and (b) statistical evaluation of score comparability, an argument can be made that cross-linguistic inferences are appropriate.**

Adapting “comparable” tests across languages

Involves

1. **Quality adaptation involving multiple steps, multiple translators, and multiple quality control checks**
2. **Statistical analysis of structural equivalence and differential item functioning**
3. **Qualitative analysis of item bias and method bias**
4. **Developing a sound validity argument for comparative inferences**

Questions that statistical techniques can help us answer

- **How “similar” are tests?**
- **How “similar” are items?**
- **How “comparable” are test scores?**

Methods for evaluating construct equivalence:

- **Differential predictive validity studies**
- **Exploratory factor analysis**
- **Confirmatory factor analysis**
- **Multidimensional scaling**
- **IRT residual analysis**
- **Differential item functioning**

Differential Predictive Validity Problem

- **A valid external criterion for differential prediction is hard to find.**

Exploratory Factor Analysis

- **Principal components analysis**
- **Common factor analysis**

Common practice is to conduct separate analysis in each language group and compare solutions.

- **Same number of factors?**
- **Items load (cluster) in the same way?**
- **Subjective comparison of solutions.**

Multigroup (simultaneous) Analyses

- **Confirmatory factor analysis (CFA) (or Structural Equation Modeling—SEM)**
- **Multidimensional scaling**
  - **Weighted MDS**
- **Unlike exploratory factor analysis, both SEM and MDS allow for simultaneous analysis across multiple groups**

Multi-group CFA Model

$$\mathbf{X}^g = \Lambda^g \xi^g + \tau^g + \delta^g$$

where $\mathbf{X}^g$ is a matrix of observed subscores for each group;

Λ^g is a matrix of factor loadings representing relationship b/w subscore & latent variable;

ξ^g represents the latent variable(s);

τ^g is a vector of intercepts i representing mean of each subscore; and

δ^g is a vector of measurement errors.

MG-CFA Analyses

- **Can evaluate different “levels” of invariance**
- **E.G., 3 hierarchical steps:**
  1. **Configural invariance**
  2. **Metric invariance**
  3. **Scalar invariance**

MG-CFA Analyses

- **Configural invariance**
  - **model (dimensionality) fits across all subgroups without constraints placed on the parameters across groups (e.g., the factor loadings and intercepts vary across groups)**

MG-CFA Analyses

- **Metric invariance**
  - **factor loadings constrained to be equal across groups (but intercepts are unconstrained).**
  - **Tests whether factor loadings are the same across groups**
  - **metric model nested w/in configural model**
  - **metric invariance tested by comparing the fit of the configural and metric models**
  - **difference in chi-square statistics ($\Delta\chi 2$) and change in CFI (ΔCFI).**
  - **ΔCFI greater than .01 may indicate a non-negligible lack of metric invariance (Cheung & Rensvold, 2002)**

MG-CFA Analyses

- **Scalar invariance**
  - **Tests whether the factor loadings and factor means (after controlling for overall proficiency) are invariant across groups.**
  - **Fit of the scalar model (i.e., with constraints placed on the factor loadings and intercepts) is compared to the fit of the metric model using the difference in chi-square statistics ($\Delta\chi 2$) and change in CFI (ΔCFI).**

Multidimensional Scaling (MDS)

- **Weighted MDS is one way to evaluate the similarity of dimensions across multiple groups**
  - **Simultaneously**
  - **Without specifying a model**
  - **(nonlinear, multigroup EFA)**

Weighted MDS model

$$d_{ijk} = \sqrt{\sum_{a=1}^{r} w_{ka}\left(x_{ia} - x_{ja}\right)^2}$$

i,*j*=items, *k*=group, *a*=dimension, *x*=coordinate, *w*=group weight on dimension

Example of using CFA & MDS to evaluate different language versions of a test

- **Sireci, Bastari, & Allalouf (1998) Psychometrics Entrance Test (PET)**
  - **Used in Israel for postsecondary admissions decisions**
  - **We looked at Hebrew and Russian versions of items from the Verbal Reasoning section of the exam.**

PET: Analysis of Construct Equivalence

- **Verbal reasoning test:**
  - **item types: (a) analogies, (b) logic, (c) reading comprehension, (d) sentence completion**
- **2 test forms, 2 language versions**
  - **Hebrew**
  - **Russian**
- **Methods**
  - **PCA**
  - **Confirmatory Factor Analysis (CFA)**
  - **Weighted MDS**

PET: CFA Results (4-Factor Model)

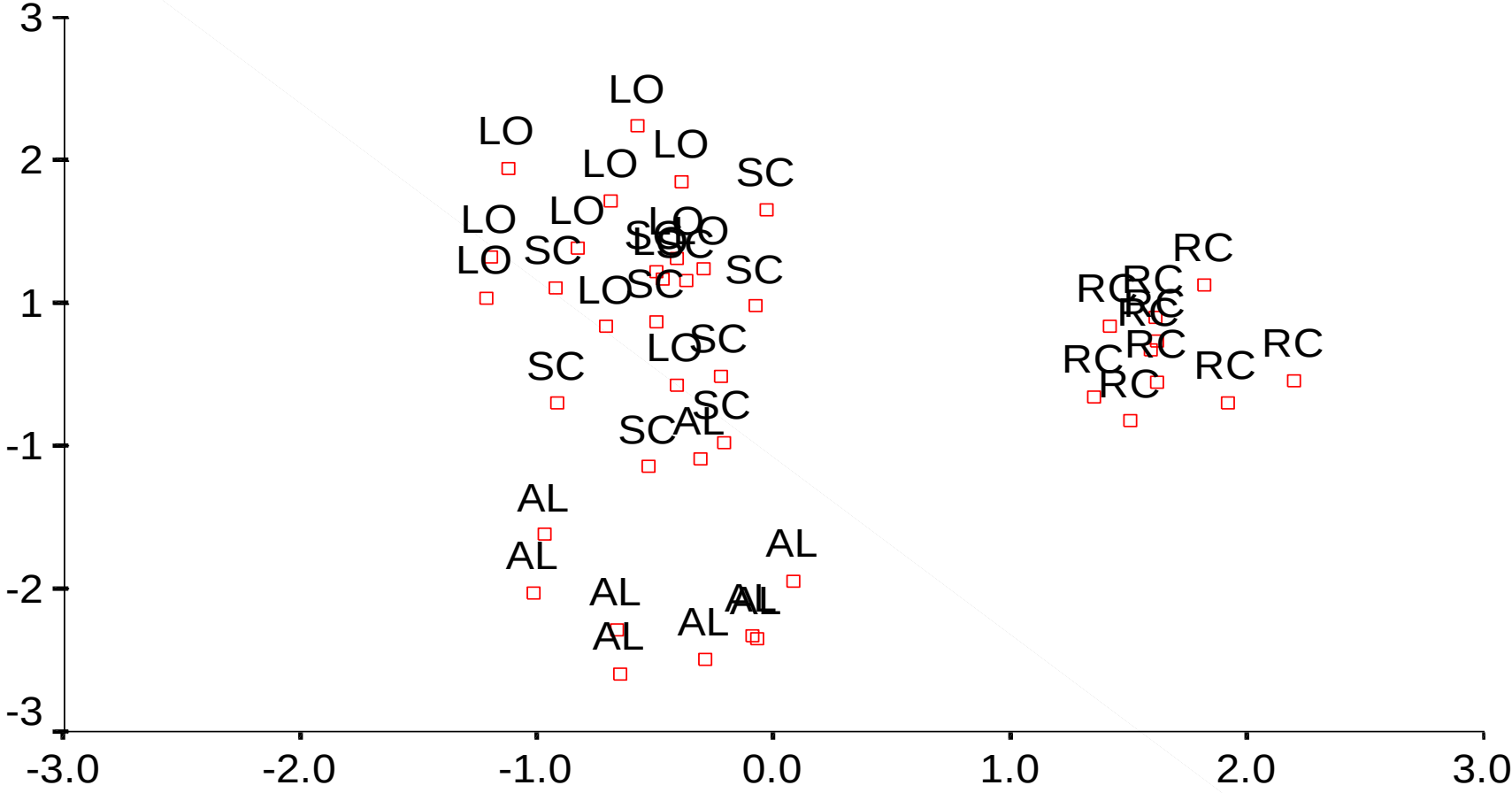
Model	GFI	RMR
Common Factor Structure	.97	.057
= Factor loadings	.96	.060
= Uniquenesses	.96	.066
= Correlations among factors	.96	.076

Next, lets look at the MDS results

- **First, we will look at the “item space,”**
- **then, we will look at the “weight space,” which contains the information regarding structural equivalence**

Figure 1

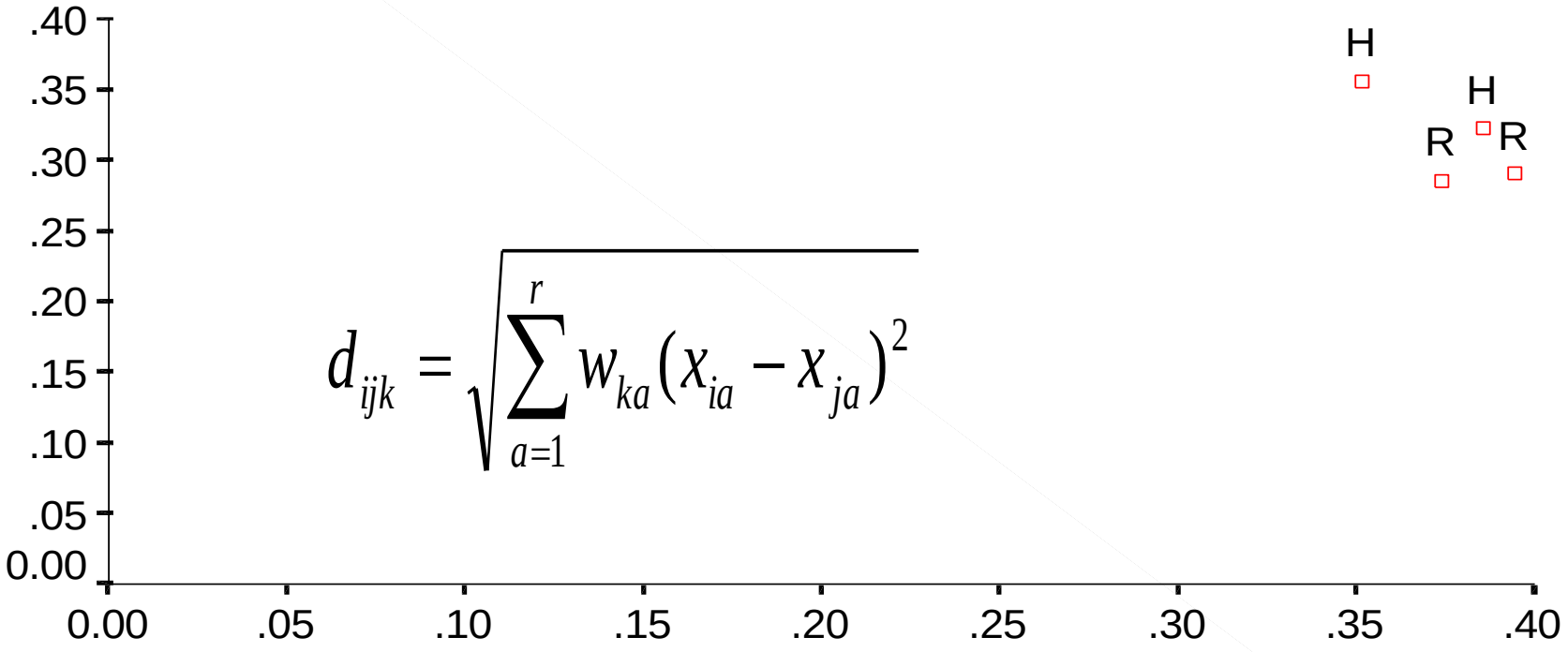
MDS Configuration of PET Verbal Items



AL=Analogy, LO=Logic,RC=Reading Compr., SC=Sentence Compl.

Figure 2

Group Weights for PET Data



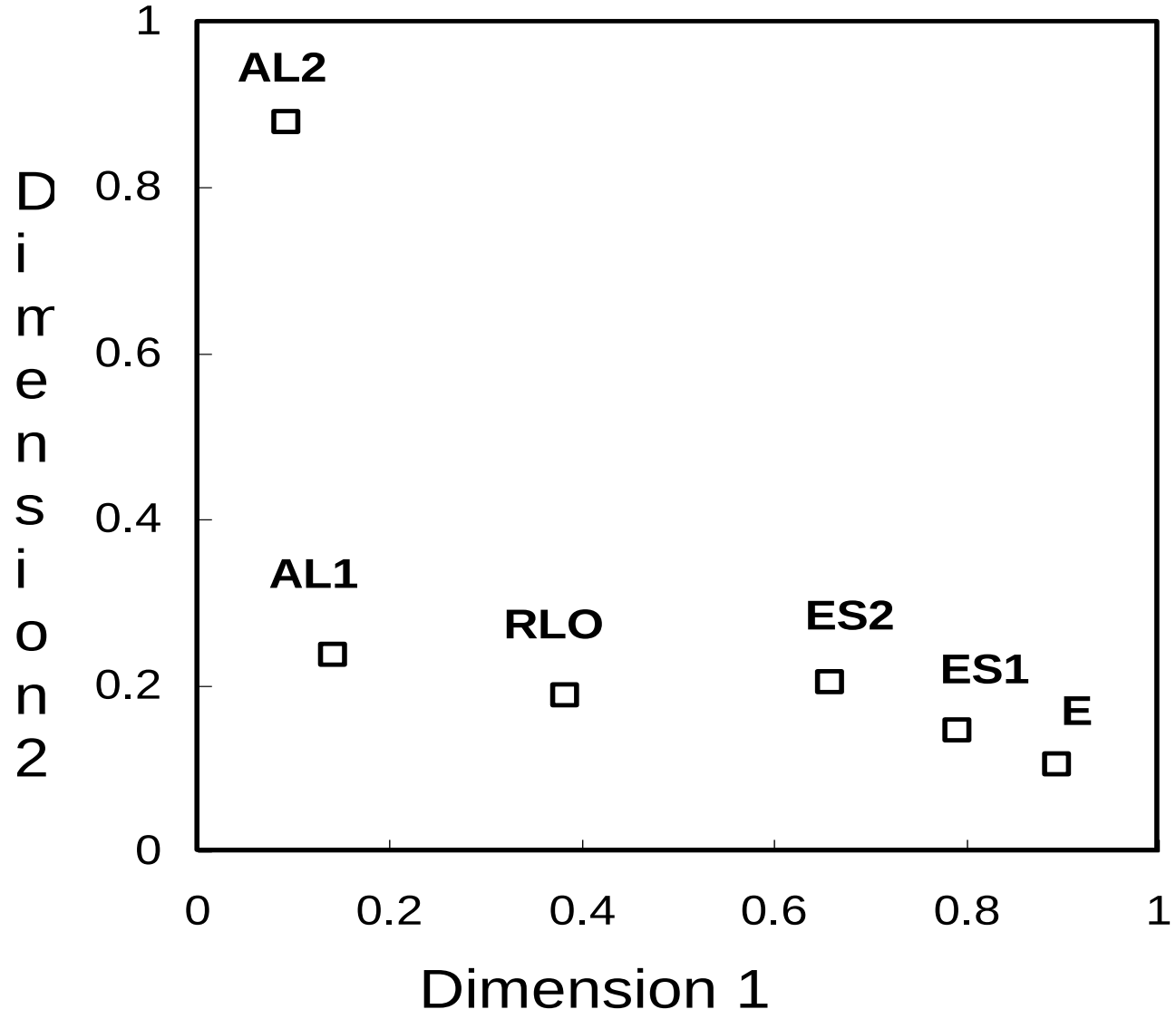
Note: H=Hebrew, R=Russian

Another MDS Example: An International Credentialing Exam (Robin et al., 2000)

- **More typical:**
  - **Large English volume**
  - **Small international volume**
- **4 Languages**
  - **English**
  - **Romance language**
  - **Two different Altaic languages**
- **International sample sizes <200.**

The next slide shows the MDS "weight space"

- **Comparing 4 language versions of the test:**
  1. **Altaic language 1**
  2. **Altaic (different) language 2**
  3. **Romance language**
  4. **English language (3 random samples, varying sample sizes to match other groups)**



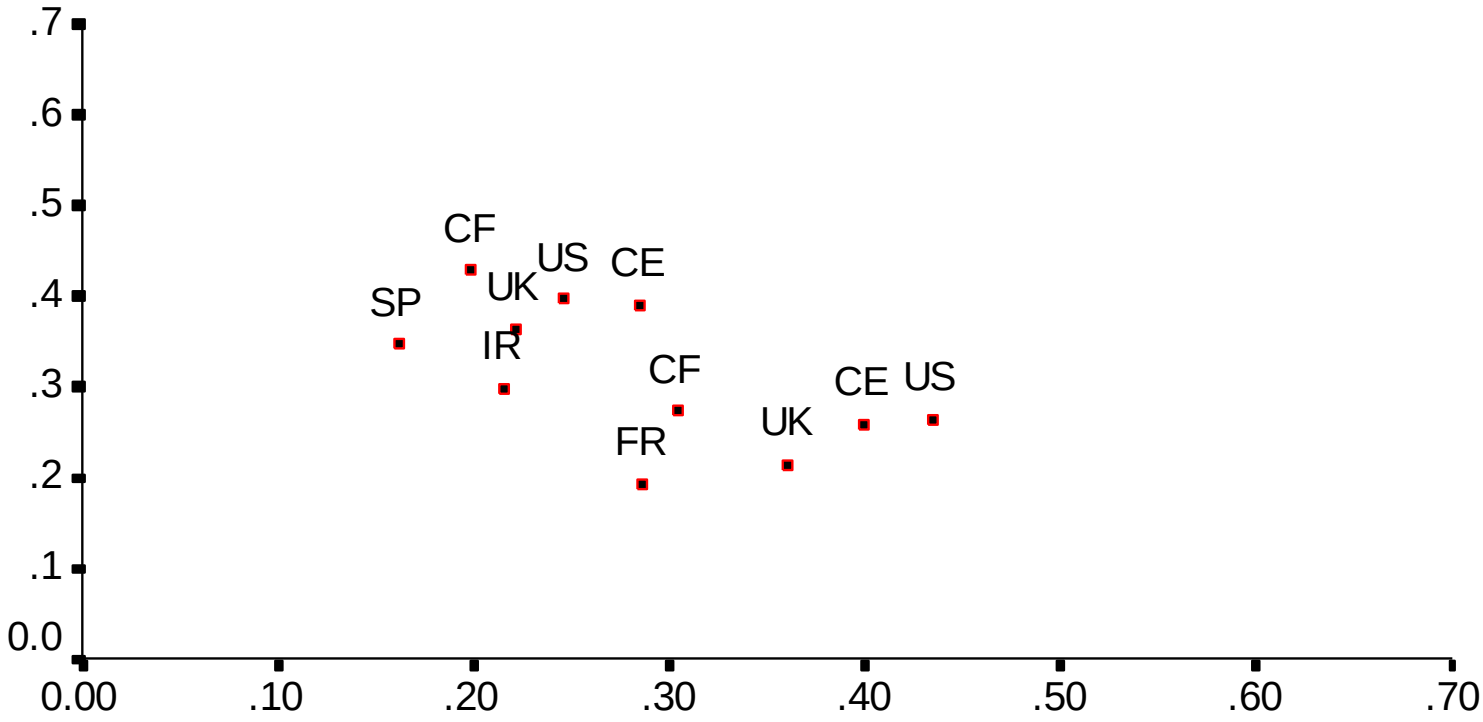
Conclusion: Different dimensions were needed to account for the structure of each language version.

Another example: Employee opinion survey (Sireci, Harter , Yang, & Bhola, 2003)

- **A very large international communications company**
- **Available in 8 languages**
- **47 different cultures**
- **Available in P&P & web formats**
- **50, 5-point Likert-type items**

Figure 3

Language Group MDS Weights



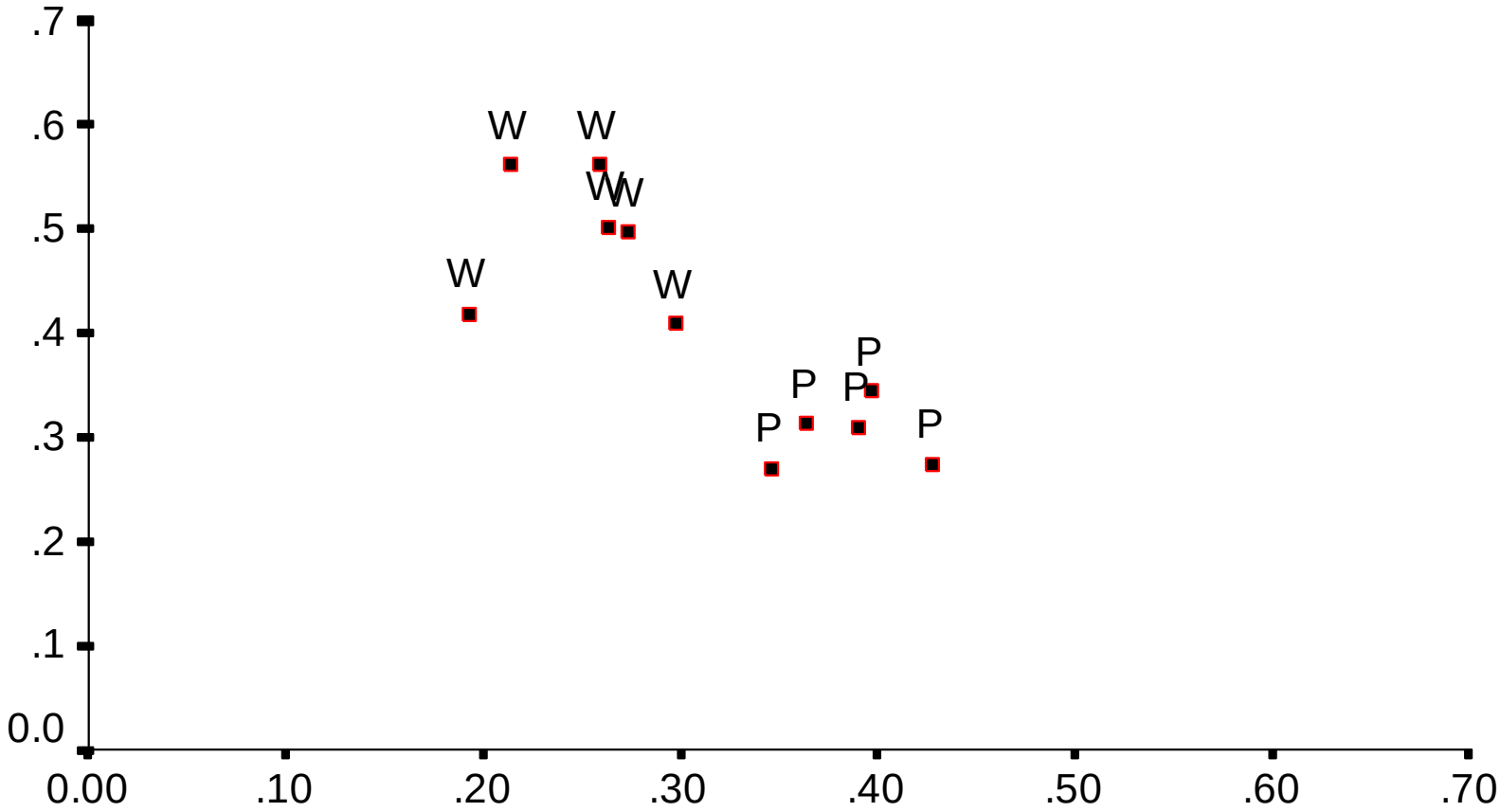
Dimension 5 (Interpersonal Relations)

CE=Can. English, CF=Can. French, FR=French, IR=Ireland (English)

UK=United Kingdom, US=United States, SP=Spanish

Figure 2

Web/Paper Group MDS Weights



P=Paper Suvey, W=Web Survey

IRT Residual Analysis

- **Fit IRT model(s) to the data for each language group.**
  - **Is fit adequate in both groups?**
  - **Are residuals (errors) small in both groups?**

  (Reise, Widaman, & Pugh, 1993)

Evaluating DIF/item bias

- **Careful translation assesses potential differences across language versions of an item.**
- **Just as item analysis catches problems item writers missed, cross-lingual DIF analyses catch translation problems.**
- **Cross-lingual DIF analyses also catch differences in cultural familiarity.**

Important Note

- **Methods for detecting DIF were not designed for studying translated/adapted items.**

- **Problem:**

(a) Cannot assume translated/adapted items are equivalent

(b) Cannot assume different language groups are equivalent

How is translation DIF different from “normal” DIF?

- **Items are NOT the “same” for studied groups.**
- **Groups cannot be considered randomly equivalent.**
- **Different groups, different items….**

How can this problem be solved?

- **It cannot be solved.**

- **However, there are (at least) 4 things that can help:**

1) Careful adaptation procedures

2) Advanced research designs

3) Aforementioned statistical analyses

4) Making certain assumptions

Systematic bias must be ruled out to justify conditioning variable in DIF analyses.

1) Careful adaptation procedures

Example: **Angoff and Cook (1988)**

- **Items in Spanish translated to English.**
- **Items in English translated to Spanish.**
- **Independent translators evaluated translations.**
- **Also, iterative DIF screening procedures.**

2) Advanced research designs

Example: **Sireci & Berberoglu (2000)**

- **Bilingual group design**
- **Two (randomly equivalent) groups of Turkish-English bilinguals took counterbalanced English/Turkish surveys.**
- **Random equivalence was tested.**
- **Polytomous (Likert) data:**
  - **Samejima's graded response model was used (IRT-LR)**

2) Advanced research designs

Sireci & Berberoglu (2000)

- **Items identified for DIF were removed from conditioning variable (theta) before making comparisons across groups**
  - **i.e., purified criterion**
- **Non-DIF items could be used to anchor scale across languages**

Other advanced research design idea:

- **Link score scales through an external criterion.**
  - **(Wainer, 1993)**
  - **Separate predictive validity studies**
  - **Logical, but has it ever been used?**
- ***Problem:*** **How to validate criterion.**

Other approaches

Use DIF screening to identify items to form link.

- **E.g., PET exam**

Allalouf, A., Rapp, J., & Stoller, R. (2009). What item types are best suited to the linking of verbal adapted tests? *International Journal of Testing, 9*, 92-107.

- **Also used by PIRLS, PISA, TIMSS**

3) Statistical analyses

- **Analysis of structural equivalence**
  - **Evaluate factorial invariance**
  - **Justify matching variable for DIF analysis**
- **“Double Linking” equating evaluation**
  - **Rapp & Allalouf (2003). Evaluating cross-lingual equating. *International Journal of Testing, 3*, 101-117.**

4) Making certain assumptions

- **Groups are randomly equivalent**
  - **Canada**
- **Anchor items are equivalent across languages (Allalouf et al., 2009)**
  - **Screened items**
  - **Non-verbal items**
- **No systematic bias**

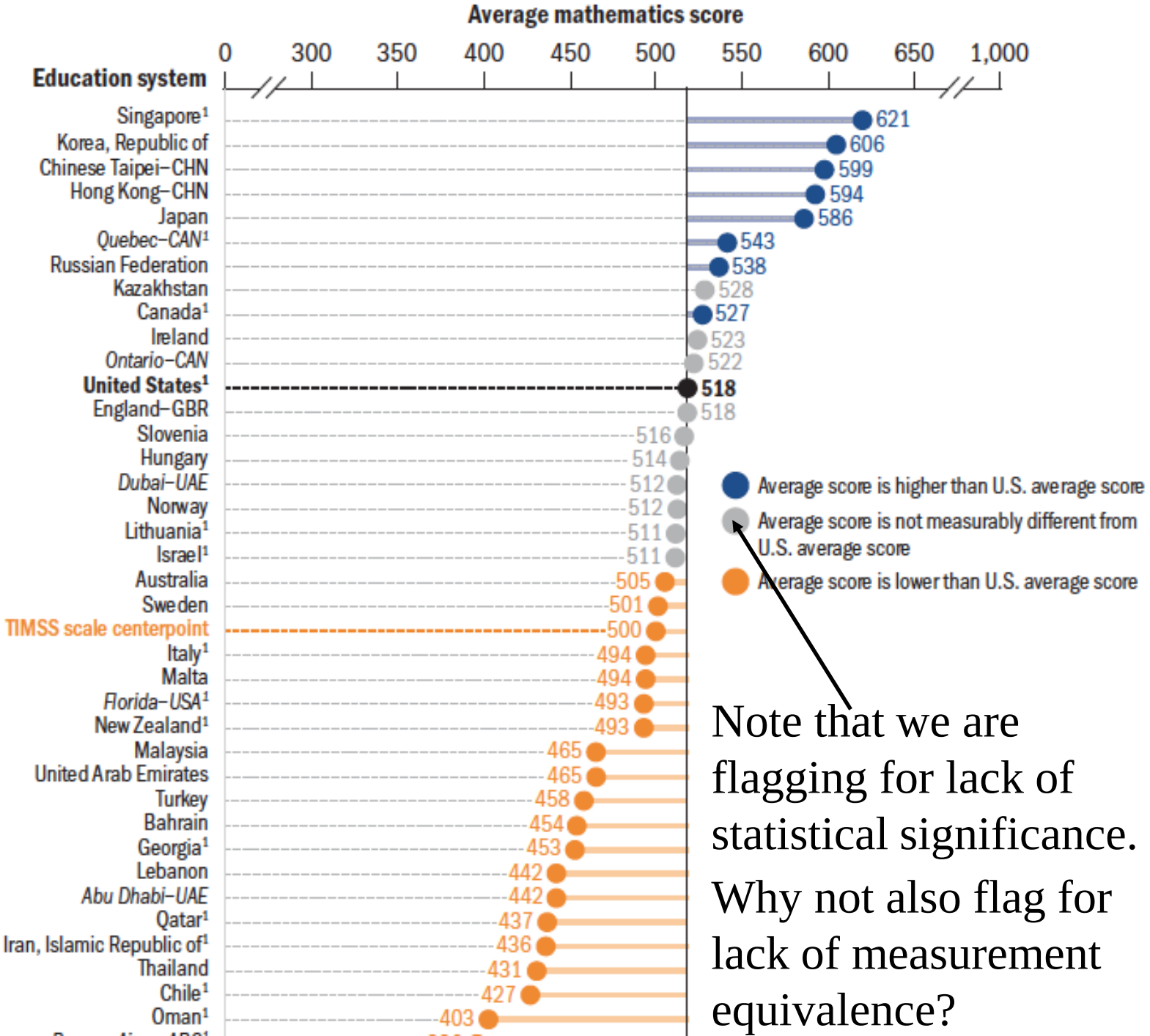
These assumptions must be defended!

Take-home messages

- **Cross-lingual assessment is hard!**
- **We can never achieve full equivalence of different-language versions of assessments.**
- **Evaluation of the validity of adaptations depends on testing purpose.**
- **For most purposes, we can achieve sufficient validity evidence.**
- **MDS is an under-used method for evaluating equivalence**
- **Many interpretations need to be qualified (“interpreted cautiously”)**

Given the problems and limitations in cross-lingual assessment

- **We need to better indicate which comparisons have evidence for validity, and which do not.**
  - **Even with the excellent translation procedures and QC controls we have heard about today and yesterday, there will still be “equivalence” or “comparability” problems.**



Note that we are flagging for lack of statistical significance.

Why not also flag for lack of measurement equivalence?

That is our challenge.

- **From qualitative and quantitative studies,**
- **Communicate what we know about comparability of cross-lingual assessment results**
- **Can we have an index of comparability?**
  - **Future research**

Next Steps/Future Research

- **What are the best ways to interpret and communicate cross-lingual inferences?**
- **Using the AERA et al *Standards* 5 sources of validity evidence to evaluate cross-lingual inferences.**
- **Using lessons learned from DIF analyses to improve future adaptations.**

Concluding remarks:

Lots of tough problems, but lots of progress.

Let’s move the field forward!

Thank you to OECD for the invitation!

And to you, for your attention.

Keep in touch

Sireci@acad.umass.edu

And see you in Montreal!

https://www.itc-conference.com