Measuring civic outcomes in large-scale surveys: Challenges and possible enhancements

Wolfram Schulz

ICCS 2016 Research Director
Content of presentation

• Measurement of learning outcomes
  – Large-scale assessments (LSA)
  – Studies of Civic and Citizenship Education (CCE)
• Measurement of cognitive outcomes with standardised tests
• Measurement of affective-behavioural outcomes and context variables
• Discussion of possible enhancements
Measurement in large-scale assessments

• No hard data but approximations:
  – Measurement error should always be taken into account
  – Many challenges in CCE-related studies common to all LSA

• Many new developments occurred in the last two decades
  – New methodologies
  – Computer technology allows faster processing
Types of measurement error

• Sampling (measurement) error
  – Representativeness of samples to infer to populations
  – No systematic bias (e.g. non-response bias)

• Instrument (measurement) error
  – Measuring reliably what we want to measure
  – Avoid systematic bias (e.g. measurement invariance across sub-groups in population)
Random error in measurement

Notice that random error doesn’t affect the average, only the variability around the average.

Systematic measurement error

Notice that systematic error does affect the average -- we call this a bias.

Examples of systematic error in measurement

- Low-achieving students not answering questionnaire (non-response bias)
- Students with low reading ability not understanding certain questions (e.g. in questionnaires)
- Students in certain contexts responding in a different way than in others (lack of measurement equivalence)
Examples of non-systematic measurement error

• Any error that is randomly distributed
  – Lower reliability but “on target”
  – Can be modelled in analysis and accounted for

• Variation in coding of open-ended questions
  – Problematic when reliability very low
Measurement of civic outcomes

• Different types of challenges/issues
  – LSA in general versus those that are CCE-specific
    • Common problems across LSA
    • Many issues not specific to CCE studies
  – National vs. international LSA of CCE
    • Specific cross-national measurement issues
## Typology of measurement challenges

<table>
<thead>
<tr>
<th></th>
<th>National assessments</th>
<th>International assessments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Common challenges for LSA</strong></td>
<td>E.g. measuring curricular aspects</td>
<td>E.g. invariance in attitude measurement</td>
</tr>
<tr>
<td><strong>Specific challenges for CCE studies</strong></td>
<td>E.g. assessing civic knowledge in absence of curriculum</td>
<td>E.g. assessing civic learning contexts across different national contexts</td>
</tr>
</tbody>
</table>
National civic assessments

• Examples
  – NAEP in United States
  – NAP-CC in Australia (Year 6 and 10)
  – SABER in Colombia (Year 5 and 9 in 2012)

• Assessing national curriculum
  – Not always in existence

• Increasingly more attention also on affective-behavioural aspects
International studies of CCE

• IEA studies
  – Civic Education studies in 1971 and 1999
  – ICCS in 2009 and 2016

• Variation in contexts
  – Challenge for instrument development and measurement

• Increasingly refined methodology
  – From one-booklet tests to rotated designs
  – More refined statistical modelling
  – Use of Item Response Theory
  – Reviews of instrument translation and adaptations
Measurement Equivalence

• Challenge in cross-cultural research
• Concerns regarding the cross-country validity of questionnaire constructs
• Methods to assess measurement equivalence:
  – Confirmatory Factor Analysis (CFA)
  – Item Response Theory (IRT)
Sources of non-equivalence

- **Language**: Linguistic non-equivalence
  - Even slight differences in wording can have effects on outcomes
  - Only obvious translation errors easily detected (e.g. item-score correlation)
- **Cultural habits**
  - E.g. “Modesty bias” (tendency in Asian countries to avoid extreme responses)
- **Differences in context** (educational system, learning practices)
  - Particularly challenging in studies on CCE
Types of measurement bias

- **Construct bias**: Construct meaningful in one context or country but not in another
- **Method bias**: Differences in responses due to cultural traits (Likert-type items)
- **Item bias**: Differential Item Functioning (DIF) due to cultural or linguistic differences
Advances in methodology

• Statistical modelling
  – Multilevel analysis taking educational structure into account
  – Structural Equation Modelling to analyse measurement structures
  – Replication (or other) methods to obtain adequate estimates of sampling variance

• Scaling
  – Item Response Theory: Described scales, sample-independent measures, equating
    • Plausible values: Better population estimates
Advances in methodology

• Instrument implementation in international studies
  – Translation verification: Independent reviews by language experts
  – National adaptation reviews: Assisting national centres with their adaptations
  – Layout verification: Ensuring that instruments are cross-nationally consistent

• Data collection
  – Quality monitoring (more systematic)
  – Computer-based assessment: Possible improvements regarding what can be assessed but still not possible everywhere
Measurement of cognitive outcomes

• Use of standardised tests
  – Criticism: No deep measurement
    • Use of open-ended items in ICCS
  – Practically the only option

• Challenges
  – Civic curricula tend not to be clearly defined as e.g. for Maths (cross-curricular, specific subjects vs. integration into other subjects etc.)
  – International studies: Diversity across national contexts
  – Definition of appropriate age group or target grade
A little bit of history...

• Early Gallup surveys since 1950s included some factual knowledge questions
• US National Electoral Surveys or Euro-barometer
  – Measurement problems due to interview assessment
  – Tend to have few items
• Establishment of national surveys (e.g. NAEP or NAP-CC)
• IEA international studies
  – CES 1971: 47 multiple-choice items (14-year-olds)
  – CIVED 1999: 38 multiple-choice items (14-year-olds) and 42 multiple-choice items for older students
  – ICCS 2009: 74 multiple-choice and six open-ended items for grade 8 students
Defining what to measure

• In national assessments
  – Specific to what students should learn according to curriculum (if there is one!)

• In international assessments
  – Defining “civic knowledge”
  – Impossible to attach strongly to national curricula (they often do not exist)
  – Need to define construct while maintaining cross-national relevance
  – Measuring civic knowledge in a similar way across countries
Ways to review/ensure appropriate measurement

• Extensive piloting and trialling
• Extensive consultation with country representatives and experts
• Monitoring how instruments are adapted/translated
• Using advanced statistical modelling to review item/scale characteristics
Measurement equivalence in cross-national testing

- The aim is to have an instrument that measures in a highly similar way across different countries/cultures/languages
  - Differential item functioning (DIF) by country indicates lack of equivalence
    - item-by-country interaction
- How can this be reviewed?
  - Field trial stage: Selecting items with less country DIF
  - Main survey stage: Reviewing this prior to inclusion for scaling
    - Identification of possible translation/adaption errors
IRT Rasch model

• Models the probability to respond correctly to an item as

\[ P_i(X_{ni} = 1) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \]

• Latent trait \( \theta \) and item parameters \( \delta \)
  – Same scale
  – Item parameter indicates item difficulty
Item-by-country interactions

• Separate calibrations for national data and pooled international sample
• Standard errors adjusted for design effects and multiple comparisons
• Graphs for each items to show variation across national sub-samples
Item-by-country interactions: Example 1
Item-by-country interactions:

Example 2
ICCS 2009: Cluster analysis of item-by-country interactions
Patterns of variation

• Clustering by language or cultural groups
  – A few exceptions: E.g. Lithuania more proximal to English-speaking countries

• Not unusual, similar results have been found for other studies (e.g. PISA reading)
  – Not specific for testing of civic knowledge
  – Some variation to be expected and perhaps even intended in cross-national studies
  – Assessment cannot be limited to those few aspects that show hardly any variation
Experience with ICCS 2009

• Described scale with different levels of civic knowledge
  – Reference to what was measured, not attached to curricular context

• Some variation in measurement characteristics (=> East Asia)

• Spanning quite a range of levels
  – Substantial differences between high- and low-performing student populations
  – Typical in international studies, may affect use of assessment in contexts with very high/low knowledge
Outlook regarding the testing of civic knowledge

• Possible to derive described scales of civic knowledge, also cross-nationally

• Assessing “deeper” knowledge aspects with other item types should be considered further
  – ICCS: Few open-ended/scored items

• Computer-based testing might add to possibilities of testing further aspects
  – Australian test delivered online in 2013
  – Not yet possible in international studies with many developing countries
Questionnaire assessments

- Civic studies tend to put more emphasis on assessing affective-behavioural variables and context information.
- Typically, questionnaires are administered to students, teachers, and school principals.
- Affective-behavioural constructs of treated as outcome variables.
  - Different from other studies like PISA, PIRLS, or TIMSS where main focus lies on achievement.
Issues with measurement

• Limits regarding the breadth of aspects that can be measured in a questionnaire
  – Typically administered after long test
  – Mainly standardised item types possible
  – All students need to understand what they are asked about

• In international studies
  – Huge variation in national contexts requires many adaptations that should be similar to original version!
  – Concerns about the comparability of responses (e.g. when using Likert-type items)
Issues with measurement

• Social desirability
  – E.g. high proportions endorsing statements on democratic values, rights for sub-groups etc.
  – Difficult to assess extent of social desirability
    • Sometimes done using separate items assumed to reflect tendency to give social desirable answers

• Finding content that is of relevance to and understood by young students
  – Concerns about choice of age group (13-14 years) for assessment in IEA studies
Target population

• Matching questionnaire content to respondents
  – Relevance of content
    • E.g. non-civic teachers should not get the impression the questions are not relevant for their work
  – Comprehension of content
    • Young adolescents should be able to understand what they are asked about
Adequacy of formats

• Clear organisation and structure
  – Limit use of repetitive item formats

• Use response categories that have clear interpretations
  – Prefer those with factual references over vague ones (not always possible)

• Have clear mappings of item material to expected constructs
  – Anticipate requirements for psychometric measurements (to achieve sufficient reliabilities)
Measuring context variables

• Gathering data on home, school or community context (e.g. ICCS)
  – Students, teachers, schools, national centres
• Reliability of measurement impacts on outcomes
  – E.g. strength of SES influences might be underestimated with weak indicators
• National contexts are quite diverse
  – E.g. school policies and practices
Cross-national comparability

• Relevance and appropriateness of questions for all national contexts
  – International options may address issues of relevance for sub-group of countries

• Ensuring appropriate adaptation of key terms in questions
  – Review introduced in ICCS 2009 (also other studies)

• Review and analysis of measurement equivalence
  – Already during instruments development (field trial)
Issues found in ICCS 2009 - 1

• Difficult to assess certain constructs
  – Democratic values: Extremely high proportions of endorsement
    • Contradicting some other findings
    • CIVED 1999: Student responses did not reflect expected patterns
  – Assessment of identity or morality issues
    • Discarded after field trial

• Many aspects seen as relevant by experts not within scope of student perceptions
  – Political institutions, policies and values
Issues found in ICCS 2009 - 2

• High proportions of missing responses in some developing countries
  – Low levels of reading literacy?
• High levels for some (civic) engagement indicators in countries with low levels of civic knowledge
  – Similar to findings in studies like PISA/TIMSS
  – Counterintuitive finding which might indicate problems with variation in measurement
  – Social desirability?
Construct validation in ICCS 2009

• Test feasibility of construct measurement
  – Analysis of item dimensionality
• Review scalability of item material
  – Measurement characteristics of items
• Inform on relationships between constructs
  – Correlations with other variables
• Review of measurement invariance
  – Analysis of cross-national psychometric characteristics (item and scale level)
Classical item review

• Relationship between item and scale
  – Item-total correlations or factor loadings

• Assessment of reliability
  – Cronbach’s alpha or other measures

• In international studies
  – Review of characteristics cross-nationally
Factor analysis

• Exploratory factor analysis
  – Exploring item dimensionality during preliminary analysis phase
  – No anticipation of factor structure

• Confirmatory factor analysis
  – Covariance-based analysis
  – Measurement model: \( x = \Lambda_x \xi + \delta \)
  – Model fit assessment
Item response theory

• Explaining probability of item response based on latent trait (here: Rasch PCM)

\[ P_{x_i}(\theta) = \frac{\exp \sum_{k=0}^{x} (\theta_n - \delta_i + \tau_{ij})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{h} (\theta_n - \delta_i + \tau_{ij})} \quad x_i = 0,1,\ldots, m_i \]

• Latent trait \( \theta \) and item/step parameters (\( \delta, \tau \))
  – Equivalent to factor analysis
  – Focus on item modelling
Review of measurement equivalence

• In ICCS 2009 we used two approaches
  – Multiple-group modelling to assess whether model holds across countries
  – IRT modelling to see consistency of parameters across countries

• Assessed at field trial stage
  – Information for item selection
Multiple-group modelling

• Multiple-group model

\[ x_g = \Lambda x_g \xi_g + \delta_g \]

• Constraining parameters across groups can be used to test hypothesis:

\[ H_\Lambda = \Lambda_1 = \Lambda_1 = \Lambda_2 = \ldots = \Lambda_g \]

• Problem with significance tests: Even smaller differences will be found statistically significant with larger sample sizes

   – When does too much variation become a problem?
Measurement Equivalence

• “Item-by-country interaction”:
  – Different measurement characteristics depending on sub-groups within a sample

• Direct estimation in facet model

\[
P_{x_i}(\theta) = \frac{\exp \sum_{j=0}^{x_i} (\theta_n - (\delta_i - \eta_c + \lambda_{ic} + \tau_{ij}))}{\sum_{h=0}^{m_i} \exp \sum_{j=0}^{h} (\theta_n - (\delta_i - \eta_c + \lambda_{ic} + \tau_{ij})))}
\]

\(x_i = 0,1,2,\ldots,m_i\)

• Review of interaction effects
Summary: Measurement equivalence

• Received more prominent attention in recent studies
  – ICCS 2009 but also in other studies
  – Should always be built in at field trial stage

• Not specific to questionnaire measurement of civic outcomes
  – However, more emphasis on questionnaires found in CCE studies

• Remaining a challenge in these studies
  – Questions about permissible variation?
Outlook: Questionnaire assessments of civic outcomes

• Development of instruments that are comparable across countries
  – Diversity of national contexts
• Targeting the respondents with appropriate instruments/content
  – Very young adolescents
  – Teachers from non-civic subjects
Outlook: Questionnaire assessments of civic outcomes

• World of civic engagement is changing
  – Questionnaire need to include new forms of civic engagement (e.g. social media)

• Emphasis on assessing more relevant forms of student engagement
  – More proximal to their school context
  – Traditional adult engagement not proximal enough

• In the future, computer-based assessment might broaden measurement possibilities
  – Problem: Not yet possible in many countries
General summary - 1

• It is possible to derive measures of civic outcomes
  – But there still remain many challenges regarding their measurement

• Importance of providing information useful for policy and practice
  – Deriving appropriate and comparable school context information remains a challenge
  – Stronger links between curricula/teaching and civic outcomes
  – Interpretation of some attitudes: E.g. is it desirable to have very high levels of trust in institutions
General summary - 2

• Further development of new item types should be considered
  – Issues with cross-national comparability and social desirability
  – General attitude statements easy to agree with

• Include more measures proximal to the students’ context
  – Changes in participation/engagement
  – Politics/political institutions tend not to be part of adolescent life
A two related questions...

• How much uniformity do we want to have in our outcomes measures?
• How much variation in measurement characteristics can be tolerated?
Thanks you!