Felianka Kaftandjieva Memorial Lecture
7th Annual EALTA Conference

Dimensionality of observations and referent generality of constructs: Some issues in modeling educational assessment data

Jan-Eric Gustafsson
University of Gothenburg
Issues of measurement in education

• General questions:
  – What is measurement and is it at all possible in the field of education?
  – How can we understand differences between qualitative and quantitative approaches?
• Specific question:
  – Issues in model based measurement
  – The requirement for unidimensionality
  – An empirical illustration, based on PIRLS
Reading Literacy

- Reading is an interactive process in which readers actively construct meaning.
- This interactive process is influenced by a large set of circumstances:
  - The reader brings a repertoire of skills, cognitive and metacognitive strategies, and background knowledge.
  - There are many different types of text which deal with different topics.
  - The context often imposes demands on the reader and influences engagement and motivation to read.
  - Social interactions about reading in communities of readers influence understanding and appreciation of texts.
Can reading literacy be measured?

- Measurement can be defined as the assignment of numbers to aspects of objects according to a rule.
- Given the complexity and context dependence of reading is it reasonable to even consider trying to measure reading literacy?
- Many say no:
  - The complexities cannot be mastered well enough, so measures are invalid and culturally biased. They therefore cause considerable damage when used, for example, in international comparisons.
  - There are fundamental theoretical and methodological problems in using measurement approaches to phenomena involving learning.
Schoultz, Säljö & Wyndham (2001)

- Schoultz et al., (2001) questioned the validity of TIMSS 1995 because it is limited to the paper-and-pencil mode of assessment.
- They also argued that performance should be seen as produced through concrete communicative practice, rather than as a consequence of students’ abilities and knowledge.
- They selected two items from the TIMSS 1995 study for scrutiny in an interview study comprising 25 Swedish grade 7 students.
- One was an optics item. It presented an illustration showing two flashlights, one with and one without a reflector, and the question was which of the two flashlights shines more light on a wall 5 meters away.
Jan and Lena each make a flashlight from identical batteries and bulbs. Lena’s flashlight contains a reflector, while Jan’s does not.

Which flashlight shines more light on a wall 5 meters away?

Explain your answer.

An open response was required, and to be scored correct the response had to include an explanation that argued that the reflector focused the light on the wall.
Schoultz, Säljö & Wyndham (2001), cont

• In the Swedish Grade 7 TIMSS sample, only 39% of the students answered the item correctly. However, in the interview study, 66% of the students gave correct answers.
• The higher performance in the interview study was to a large extent due to the scaffolding provided by the interviewer in a Socratic dialogue.
• Schoultz et al. concluded that: “Knowing is in context and relative to circumstance. This would seem an important premise to keep in mind when discussing the outcomes of psychometric exercises.”
• This may seem as serious criticism not only of the validity of the TIMSS study, but also of results from attempts to make assessments of knowledge with paper and pencil tests generally. However, this study does not really address the issue of validity of the TIMSS assessment.
Schoultz, Säljö & Wyndham (2001), cont

- Schoultz et al. start from the assumption that performance differences between different contexts are absolute, and that the higher performance in the interview situation is evidence of a higher level of knowledge and conceptual insight, i.e., evidence of higher student ability.
- In the measurement approach used in TIMSS, performance differences are seen as relative, because the observed performance level is interpreted as being determined not only by student ability but also by the difficulty of the item. According to this view, one possible interpretation is that the item is easier in the interview situation than in the paper and pencil situation.
Assumptions and metaphors

• Is the absolute or relative view of performance differences correct? None, and both! Complex phenomena cannot be described unless we see them from a perspective. One way to capture different perspectives is to describe them in terms of metaphors.
• Sfard (1998) proposed that learning may be described in terms of either an acquisition metaphor or a participation metaphor.
  – The acquisition metaphor views knowledge as an acquired commodity, so learning is a process of acquisition with individual ownership of knowledge as a result.
  – The participation metaphor views learning as taking part in a collective and communicative process.
  – Both metaphors are limited, and we should not only choose one of them.
Another metaphor: weather and climate

• Weather affects our daily lives, how we dress, what we do and talk about. We may adapt to weather but there is not much we can do about it. In the short run we can predict weather, but beyond a week or so weather is unpredictable.

• Climate is generalized weather over a longer period of time. We experience weather, and through aggregating these experiences, we get a sense of climate. In a more precise manner scientists define climate as aggregate aspects of weather, using indicators such as mean temperature and mean rainfall. Thus, climate is an abstraction.

• While weather is unpredictable and chaotic, climate and climate changes are stable phenomena, which we can understand theoretically and for which empirically based models may be constructed, that predict long-term development.

• Climate is a social construction, and research on climate is based on a highly developed technology of devices for generating data, on agreed-upon definitions, and analytical models. But the fundamental idea is to aggregate multiple observations of different aspects of weather.

• In the same manner quantitative research in education is based on aggregation of observations of different aspects of phenomena of teaching and learning.
Is aggregation good or bad?

• Yanchar and Williams (2006) argued that: “… data aggregation and accompanying statistical tests often hide qualitative patterns and lead to excessively abstract or artificial conclusions …; statistical indices are often used as facile substitutes for careful interpretation and human judgment … patterns in aggregate data are erroneously used to make inferences about the structure of psychological processes in individuals …”
• But the argument can also be turned around, and it can be argued that in order to see the general aspects (e. g., the climate) it is necessary to get rid of the specifics (e. g., the weather). Seen from this perspective methods which conceal context-dependent variation have strengths, rather than disadvantages, when the purpose is to investigate general patterns and relations.
• Thus, aggregation may be both good and bad, depending upon the purpose of the research.
• Aggregation can also be done in ways that are more or less useful.
The quantitative/qualitative dichotomy

• Much methodological debate starts from a dichotomy between quantitative and qualitative methods, and associated distinctions between objective/subjective, positivistic/hermeneutic, nomothetic/ideographic, and bad/good.
• Ercikan and Roth (2006) argued that the quantitative and qualitative dichotomy is fallacious:
  – Quantitative research is typically based on qualitative distinctions in data generation and in conceptualisation
  – Much qualitative research aims at, and does achieve, generalizations.
• They proposed that different forms of research should instead be put on a continuous scale that goes from the lived experience of people on one end (low-level inference) to idealized patterns of human experience on the other (high-level inference).
  – Low-level inference research is characterized by contingency, particularity, being affected by the context, and concretization,
  – High-level inference research is characterized by standardization, universality, distance, and abstraction.
Measurement

- Aggregation of observations is used to create measures of abstract constructs, such as reading literacy skill, self-efficacy, introversion, intelligence, and quality of teaching.
- With aggregation over observations stability and consistency (reliability) is gained, according to the general principle that combination of many fallible observations from different contexts and contents yields a less fallible aggregated score.
- However, aggregation over contexts and contents may cause the meaning of the aggregate to become unclear or get lost entirely.
- If we do not measure the intended constructs we have problems of construct validity. There are two major sources of threats against construct validity:
  - Construct irrelevant variance (the measure is influenced by irrelevant factors, such as reading ability or social desirability).
  - Construct underrepresentation (the measure does not fully cover the intended construct, perhaps because the data collection methods impose restrictions on the type of observations that may be obtained).
The Rasch model

- According to the Rasch model the probability of a correct response to a test item is a function of the ability of the person, and of the difficulty of the item.
  - As the ability of the person increases, the probability of a correct answer increases.
  - As the difficulty of the item increases, the probability of a correct answer decreases.
- It is possible to estimate the difficulty of items independently of the ability of persons, and to estimate the ability of persons independently of the difficulty of the items.
- It follows from the Rasch model that the total score on a set of items includes all the information about the person’s ability. Thus, when the Rasch model is applicable, we can make sensible aggregations of scores on different items.
Assumptions of the Rasch model

• The Rasch model is based on two fundamental assumptions:
  – Homogeneity of item characteristic curves: A single parameter (item difficulty) is sufficient to represent the statistical behaviour of the item.
  – Unidimensionality: all items measure one and the same ability
• When these assumptions are violated, the attractive properties of the Rasch model are threatened.
• Test of fit of data to the model frequently indicate poor fit of single items and groups of items, and often the unidimensionality assumption is violated. Different strategies are possible to deal with this:
  – Delete the poor-fitting items from the test.
  – Split the items into subsets, and create two or more scales.
  – Accept the poor fit, and continue as if fit was good.
An example

• Gustafsson and Lindblad (1978) used the Rasch model to analyze a test of English grammar for Swedish students.
• There was poor fit of the model to data, which was primarily due to poor fit of a set of items measuring knowledge of irregular verbs.
• In separate analyses, good fit was found for the irregular verb items, as well as for the other items, after some poorly constructed items had been excluded.
• Exclusion of the irregular verb items would have caused unacceptable construct underrepresentation, but through creating separate scales the items were retained.
A dilemma

• Violations of the assumption of unidimensionality can be solved by creating multiple scales.
• However, this solution creates new problems because we may have to splinter our domain of measurement into a large number of narrow scales, which may be impractical and theoretically inproductive.
The standard view of measurement and its problems

• The standard view of measurement:
  – The phenomenon we want to capture can be described in terms of a set of correlated dimensions, which all are unidimensional.
  – There is no other relation among the dimensions than correlations (i.e., no causal relations, or sub- or superordination).

• Problems with the standard view:
  – Analytical techniques such as regression analysis assume each independent variable to measure one ’thing’ only. However, the dependent variable is understood in terms of multiple sources of variance. Thus, while the independent variables are viewed as unitary, the dependent variable is viewed as complex.
  – Some constructs are broad and encompass a wide range of phenomena (e.g., reading literacy), while other constructs are narrow and encompass a more limited range of phenomena (e.g., knowledge of irregular verbs). Along with the unidimensionality requirement this implies a focus on constructs with narrow referent generality, and typically it implies that broad constructs are splintered into more and more narrow constructs.
Hierarchical factor-analytic models

• The Rasch model belongs to a large class of so-called latent variable models in which latent variables are related to manifest variables. Factor-analytic models is another example.
• Hierarchical factor-analytic models allow identification of both broad and narrow sources of variance. In so-called nested-factor models, some factors have relations to many manifest variables, while others have relations to few manifest variables.
• From the hierarchical model it follows that observed variables are complex, and reflect dimensions of different degrees of generality.
A hypothetical example

In this example performance on a reading passage is influenced by a general reading comprehension factor (ReadComp), a speed factor (ReadSpeed) affecting some passages, and a constructed response factor (ConstResp) which also affects some passages.
A real example: The 1991 Reading Literacy Study (RL) and the 2001 Progress in International Reading Literacy Study (PIRLS)

• In 1990/1991 the International Association for the Evaluation of Educational Achievement (IEA) conducted an international comparative study of reading literacy, and in 2001 the IEA started the PIRLS study, which is to be conducted every fifth year.
• In both these studies the main aim is to measure reading literacy and both start from a functional definition of reading literacy: “... the ability to understand and use those written language forms required by society and/or valued by the individual.”
• However, frameworks defined for the assessments were quite different, specifying partly different types of texts, and using items with different formats. It may thus be asked what different aspects of reading literacy are being captured by the RL and PIRLS studies.
Instruments in RL and PIRLS

• The RL Instruments
  – Four narrative prose, five expository prose, and six documents text passages.
  – Two booklets: Booklet A (35 minutes) and Booklet B (40 minutes), each with a mixture of text types.
  – 66 multiple-choice items measuring different aspects of reading comprehension.
  – Every student took every task.

• The PIRLS instruments
  – Eight textblocks, four literary texts and four Information texts. The blocks were rotated across 10 different booklets, two blocks in each.
  – Every student in the PIRLS sample took one of the booklets (2 x 40 minutes were allowed). Total testing time in PIRLS thus was 320 minutes.
  – About a dozen items for each textblock, half the items being constructed response items, while the other half was multiple-choice items.
  – The items measured different aspects of reading comprehension.
  – One overall reading literacy scale was created using IRT, along with two sub-scales,
Differences between RL and PIRLS instruments

- A large number of Documents texts in RL.
- In RL all items were multiple-choice items, while half the items in PIRLS were constructed-response items.
- In PIRLS a subset of the items focus upon the textual representation itself.
- The texts included in PIRLS in general were longer than those in RL.
The Swedish PIRLS 2001 assessment design

- For PIRLS the regular rotation of booklets within classes was used.
- Every other class got RL Booklet A and every other class got RL Booklet B.
- This created a design with planned missing data, but with sufficient information to allow use of multivariate modelling procedures taking missing data into account.
- A sequence of nested-factor models was fitted to 15 block scores from RL and to 16 half-block scores from PIRLS.
Models for RL and PIRLS

- ReadComp_PIRLS to ReadComp_RL: 0.90
- ConstrResp to ReadComp_PIRLS: 0.93
- ConstrResp to ReadComp_PIRLS: 0.96
- ConstrResp to ReadComp_PIRLS: 0.99

Documents

EndOfTest

ConstrResp

Faculty of Education
Unidimensionality and aggregation

- The general factor dominates both tests.
- For the RL sum, the factors account for variance as follows:
  - General reading comprehension factor 79 %
  - End of Test factor 11 %
  - Documents factor 2 %
  - Block factors 3 %
  - Error 4 %
- The principle of aggregation:
  - Aggregation causes the general factor to account for a larger proportion of variance in the sum of scores than it does in each item.
  - Each item is complex, but aggregate scores may be unidimensional.
Conclusions

• Constructs with broad referent generality are practically important constructs that also need to be understood theoretically. Such constructs capture a clustered set of attributes that may share a common causal mechanism or there may be mutual causation among attributes.

• In both classical and modern test theory the way unidimensionality is conceptualized and analyzed tend to imply a focus on constructs with narrow referent generality rather than on constructs with broad referent generality.

• Aggregation over broad domains of performance may be a way to approximate unidimensionality so that robust use of the Rasch model and other IRT models may still be possible.