The dual nature of tests: research instrument, and object of research

Dr Claudia Harsch
CAL, University of Warwick

Overview

- Tests as Research Instruments
  - Purposes – when to use which one?
  - Quality criteria – how to know you've chosen a good one which is fit for purpose?
- Tests as Objects of Research
  - Qualitative and quantitative properties
  - What should a good test look like?
- Let's start with your research…

Research design

- Take a minute or so to outline your research design – is it…
  - qualitative, quantitative, mixed-methods?
  - pre-, quasi-, or experimental?
- What hypotheses and research questions do you have?
- How will you answer them?
- Do tests play a role? Which one?

Research design

- What are the characteristics and likely problems with the three experimental approaches?
  - pre-experimental
  - quasi-experimental
  - experimental

What we can use tests for…

- Compare measures before and after an intervention (e.g. for effectiveness)
- Compare measures between different groups (e.g. to control or examine background variables and their influence)
- Confirm or develop hypotheses
- Examine the effect of certain variables on an outcome (looking at dependent, independent and moderating variables)
- …

How can we analyse our test-based data?

- Relationship between variables:
  - Correlation
- Effects of variables on dependent variables :
  - Regression analysis
- Comparing groups, variables:
  - Frequencies between two groups: Chi-square
  - Means between two groups or points in time: t-tests
  - Comparing more than two groups, variables: ANOVA

How can we analyse our test-based data?

- Reduce variables to common underlying factors:
  - Factor analysis (exploratory or confirmatory)
- Examine causal relationships between variables:
  - Structural equation modelling
- NB: Each of the methods of analysis comes with its own assumptions and restrictions!
- You need to know about them, and make sure your data set complies with these!

What do we need to bear in mind?

- Pre-/post-test designs
  - Test equivalence?
  - Construct
  - Difficulties
  - Learning effects
- Designs to examine effects of intervention
  - How to control other factors besides intervention?

What else do we need to bear in mind?

- Do you have the time, resources and *assessment literacy* to develop, pilot, analyse, revise, validate and evaluate your test?
  - Is anybody here developing their own test?
- Is there an existing instrument which you could adapt for your study?
  - Who is using an existing test?

How to select a test which is fit for your purpose

- Is the test and research on it published?
- Do the test and your study have a comparable purpose?
- Does the test aim at your group of participants?
- Are the construct(s) of the test and your study comparable?
- Are the test items adequate for your study and participants?
- Could the use of the test in your study have unintended consequences?
- Can you get the test authors' permission?

When is a test a 'good' test?

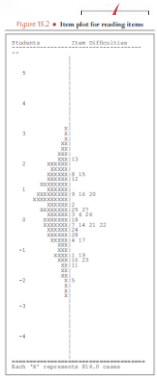
Look at published test and research on its quality:

- Aims, target group, purpose
- Construct, specifications
- Test items and assessment criteria
- Reliability, item properties
- Validity
- Impact studies

Properties of a good test I

- Reliability:
  - Does the test yield reliable result, independent from test sitting, marker, context conditions...?
  - Test-retest, parallel forms, split-half, rater reliabilities...
  - Cronbach's alpha as indicator of lower margin of reliability
- Validity:
  - Does the test measure what is claims to measure?
  - Qualitative studies into cognitive processes
  - Dimensionality analyses
  - Examining difficulty-determining properties, predicting dificulties, ...

Properties of a good test II



- Item properties: CCT – valid only for tested sample!
  - Facility value (% correct), standard deviation
  - Distribution of test takers, standard error of measurement
  - Reliability, Cronbach's alpha
  - Discrimination index (item-total correlation)
  - Distractor analyses for items with distractors
  - Analysis of item bias
- IRT: valid for population represented by sample
  - Probabilistic models to estimate item difficulty and student proficiency on the same scale
  - Check item properties (infit, outfit, bias)

Validation studies

- Use validated instruments as point of comparison (e.g. Bae & Lee 2011, Fitzpatrick & Clenton 2010)
  - E.g. Correlation, regression, factor analyses
- Use item-difficulty modelling to predict difficulties and test construct hypotheses (e.g. Harsch & Hartig, 2011)
- Use qualitative studies, e.g.
  - Cognitive processes when solving items via think-aloud (e.g. Rossa 2012)
  - Examine test takers' perspective and perception (e.g. Crisp et al., 2008)
  - Use expert evaluation, e.g. of test items or of rating scales (e.g. Harsch & Rupp 2011, Harsch & Martin 2012a, b)

Impact studies

- Overview of current research approaches in Taylor & Weir (2009)
- Research methods in Cheng et al. (2004)
- Examples for such studies: Alderson & Hamp-Lyons (1996), Green (2007), Wall (2005)
  - Baseline study needed to be able to examine impact of a newly introduced test on its social or educational context

Further Qualitative Analyses

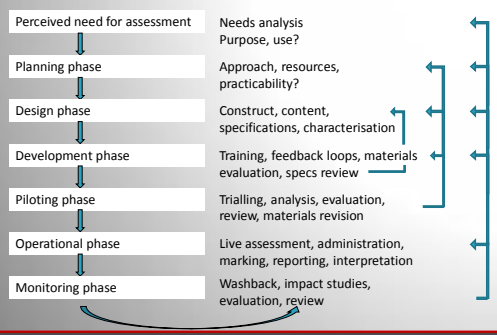
- Content analyses: Expert judgements
- Discourse analyses
  - Test discourse ≠ authentic communication
  - Effects of tasks, test takers, raters, interviewers
  - Test comparison
- Introspection
  - Focus on test takers, interlocutors, raters
  - Concurrent (think-aloud) or retrospective
- Context analyses
  - Ethnographic methods, e.g. interview, observation, questionnaires

Further Quantitative Analyses

- Comparisons, looking for similarities
  - Compare different tests / versions (correlation, regression, factor analysis)
  - Rating reliability and validity (IRT e.g. FACETS, G-theory)
  - Analysis of dimensionsionalities (factor analysis)
- Analysing differences in variance
  - ANOVA, t-tests, Chi-square
  - Item bias: DIF analyses
- Predicting effects, examining causality
  - Regression analyses
  - SEM

In case you are working on developing a test instrument, let's look at the necessary steps ....

Cyclical Model of development process (Milanovic, 2002)



Training in assessment literacy, test development and analysis

There are courses and workshops on offer, e.g.

- MA ELT with specialism in Testing and assessment here at Warwick, CAL
- EALTA pre-conference workshops
- LTRC workshops
- Workshops offered by ALTE
- Summer schools, e.g. EALTA or University of Lancaster

Checklist for evaluating test instruments for research purposes

- Based on the EALTA guide for good practice http://www.ealta.eu.org/guidelines.htm
- Handout

References

- Alderson, J. C. & Hamp-Lyons, L. (1996), TOEFL preparation courses: a study of washback. *Language Testing 13:3, 280–297.*
- Bae, J. & Lee, Y. (2011). The validation of parallel test forms: 'Mountain' and 'beach' picture series for assessment of language skills. *Language Testing 28:2, 155–177.*
- Cheng, L., Watanabe, Y. & Curtis, A. (2004). *Washback in Language Testing: Research Contexts and Methods. Mahwah, NJ: Erlbaum.*
- Cohen, L., Manion, L. & Morrison, K. (2011). *Research Methods in Education* (7. Aufl.). London and New York: Routledge.
- Creswell, J. & Plano Clark, V. (2007). *Designing and Conducting Mixed Methods Research*. London: Sage.
- Crisp, V., Sweiry, E., Ahmed, A. & Pollitt, A. (2008). Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions. *Educational Research* 50:1, 95–115.
- Fitzpatrick, T. & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing 27:4, 537–554.*
- Green, A. (2007). Washback to learning outcomes: a comparative study of IELTS Preparation and University presessional language courses. *Assessment in Education 14:1,* 75–97.
- Harsch, C. (2012). Der Einsatz von Sprachtests in der Fremdsprachenforschung: Tests als Untersuchungsgegenstand und Forschungsinstrument. In S. Doff (Ed.), *Empirisch basierte Methoden in der Fremdsprachenforschung: Eine Einführung* (pp. 150-183). Tübingen: Narr.

References

- Harsch, C. & Martin , G. (2012b). Using descriptors as the basis for analytic ratings – improving level-specific writing assessment. In: *Assessment in Education.* Pre-publication online, doi: 10.1080/0969594X.2012.742422.
- Harsch, C. & Martin , G. (2012a). Piloting a new rating scale: rater training and scale revision combined. In: *Assessing Writing* 17, 228–250.
- Harsch, C. & Hartig , J. (11/2012). *Effects of selected item characteristics on difficulty and dimensionality of reading and listening comprehension tests*. Paper presented at the Language Testing Forum, Bristol, UK.
- Milanovic, M. (2002): *Common European Framework of Reference for Languages: Learning, Teaching, Assessment - Language examining and test development.* Online available at: http://www.coe.int/T/DG4/Portfolio/documents/Guide%20October%202002%20revised%20version1.doc
- Porte, G. (2010). *Appraising Research in Second Language Learning. A practical guide to critical analysis of quantitative research (2. Aufl.). Amsterdam: Benjamins.*
- Rossa, H. (2012). *Zur Validität der Messung sprachlicher Kompetenzen: Eine Untersuchung der Prozessebene der Aufgabenbearbeitung am Beispiel von Testaufgaben zum fremdsprachlichen Hörverstehen. Frankfurt/Main: Lang.*
- Taylor, L. & Weir, C. (eds) (2009). *Language Testing Matters: Investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge Conference, April 2008. Cambridge: Cambridge ESOL and Cambridge University Press.*
- Wall, D.(2005). *The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory. Cambridge: Cambridge ESOL and* Cambridge University Press.

Further reading

- Bachmann, L. & Palmer, A. (2010). *Language Assessment in Practice*. Oxford: OUP.
- Bachmann, L. (2004). *Statistical Analyses For Language Assessment*. CUP.
- Douglas, D. (2010). *Understanding Language Testing*, London: Hodder.
- Fulcher, Glenn & Fred Davidson, (2007). *Language testing and assessment. An advanced resource book*. London: Routledge.
- Weir (2005). *Language Testing and Validation*. Oxford: Palgrave.

The following books are recommended for assessing different language skills:

- Alderson, J.C. (2000). *Assessing Reading*. Cambridge: CUP.
- Buck, G. (2001). *Assessing Listening*. Cambridge: CUP.
- Chapelle, C. & Douglas, D. (2006). *Assessing Language through Computer Technology.* CUP.
- Luoma, S. (2005). *Assessing Speaking*. Cambridge: CUP.
- Weigle, S.C. (2001). *Assessing Writing*. Cambridge: CUP.

URLs of assessment associations:

- ALTE: http://www.alte.org/
- EALTA: http://www.ealta.eu.org/; http://www.ealta.eu.org/guidelines.htm
- ILTA: http://www.iltaonline.com/