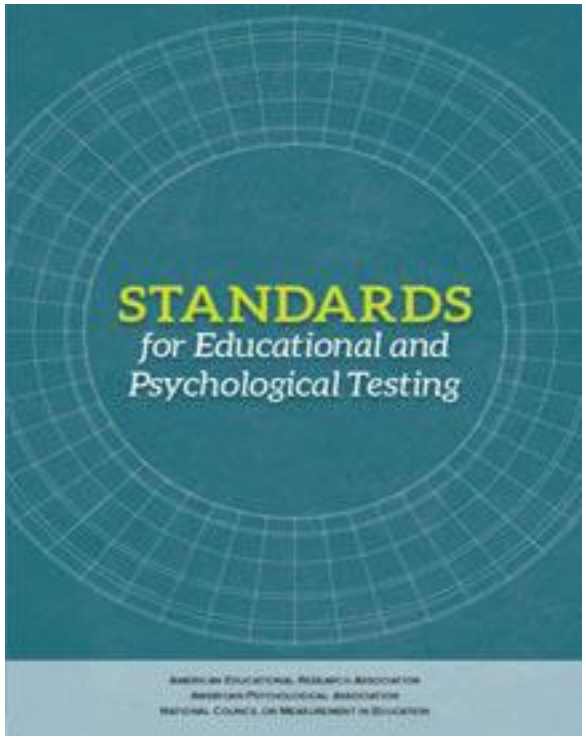
Nederlands Instituut van Psychologen NIP

Dutch Committee on Tests and Testing (COTAN): Current trends in testing

Rob Meijer


STANDARDS
for Educational and
Psychological Testing



The *Standards for Educational and Psychological Testing* has been produced through a long-standing collaboration of three associations: the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME).

In the past 15 years, important developments have occurred in the field of testing,
requiring significant revision to the *Testing Standards*. Five areas, in particular, receive attention in the 2014 revision:

Examining accountability issues associated with the uses of tests in educational policy
Broadening the concept of accessibility of tests for all examinees
Representing more comprehensively the role of tests in the workplace
Taking into account the expanding role of technology in testing
Improving the structure of the book for better communication of the standards

Outline

- **Introduction COTAN**
- **http://www.psynip.nl**
- **Dutch Rating System for Test Quality**
- **Rating procedure**
- **Trends and developments**
  - Computer-based and Internet-delivered testing
  - Short scales
  - Continuous norming
  - Fairness

Intro COTAN (1)

Dutch Committee on Tests and Testing in the Netherlands

- **One of the several committees of the Dutch Association of Psychologists (NIP)**
- **'Mission':**

  improving test use and test quality by informing test constructors, test users, and test publishers about the availability, the content, and the quality of tests

  - **to audit the quality of Dutch psychological instruments**
  - **to raise standards in the use of psychological tests**

Intro COTAN (2)

Dutch Committee on Tests and Testing in the Netherlands

- **To audit the quality of Dutch psychological instruments**
  - instruments that are applied in settings like schools, human resource management, (mental) health care services
  - *Dutch Rating System for Test Quality*
  - *Documentation of Tests and Test Research*

    digital database containing a description of all tests available in the Netherlands and contains excerpts of research conducted with these tests

    www.cotandocumentatie.nl

Intro COTAN (3)

Dutch Committee on Tests and Testing in the Netherlands

- **To raise standards in the use of psychological tests**
  - *General Standardised Guideline for the Use of Tests*
    describes guidelines for the professional use of tests for diagnostic purposes

Composition of COTAN

- **Members:**
  - 1 representative from each university
  - NIP committee members
    - from the sectors Work & Organization, Health care, and Youth
  - Members from leading companies
    - like KLM and Cito

- **At this moment, 19 COTAN members**
- **There is also a pool of external test raters**
  - Different professionals; researchers as well as practitioners

- **Volunteers!**

Dutch Rating System for Test Quality (1)

The rating system provides grades for seven criteria:

1. **Theoretical basis of the test**
2. **Quality of the test materials**
3. **Comprehensiveness of the test manual**
4. **Norms**
5. **Reliability**
6. **Construct validity**
7. **Criterion validity**

Dutch Rating System for Test Quality (2)

1. **Theoretical basis of the test**

   theoretical basis and logic of the test development procedure

   - Definition and operationalization of construct; description of groups for which test is meant; application of the test; theoretical elaboration of test construction process

2. **Quality of the test materials**

   for both P&P version as computer-based version

   - standardization of the test content; objectivity of scoring system; presence of unnecessary culture-bound words or content that may be offensive for specific groups; instructions for test taker

Dutch Rating System for Test Quality (3)

3. Comprehensiveness of the test manual

comprehensiveness of the information in the manual to enable well-founded and responsible use of the test by the test user

- Completeness of instructions for successful test administration, inclusion of case descriptions, availability of indications for test-score interpretation, statements on user qualifications

4. Norms

norm-, domain-, and criterion-referenced interpretation

- Availability of norms, size, representativeness, year or period of data collection

Dutch Rating System for Test Quality (4)

5. Reliability

different requirements for tests intended for making important decisions and less important decisions w.r.t. level of reliability

- quality of reliability research design and completeness of supplied information
- use of other methods than coefficient alpha are allowed and encouraged, like Guttman's lambda2 and greatest lower bound

6. Construct validity

Different types of research in support of construct validity: research on dimensionality of item scores, psychometric quality of items, invariance of the factor structure and possible bias, convergent and discriminant validity, differences between relevant groups

- quality of research design and completeness of supplied information

Dutch Rating System for Test Quality (5)

7. Criterion validity

evidence should demonstrate that a test score is a good predictor of non-test behavior or outcome criteria. Prediction can focus on the past (retrospective validity), the same moment in time (concurrent validity), or on the future (predictive validity)

- quality of research design and completeness of supplied information

Rating procedure

- Free of charge for test author and/or test publisher
- Two anonymous raters
  - Comparable to peer reviews in research
  - COTAN member and external rater
- Differences in ratings will be discussed
- Editor combines ratings into one description and summarizes grades for seven criteria
- Test author/publisher can lodge an objection once
- Due to availability of the raters, re-assessment of an instrument after at least one year

Purpose of ratings

1. to inform test users about the quality of available instruments. This information can be helpful in choosing the proper instruments for their setting.
2. to supply feedback to test developers about the quality of their products. For them the rating system can serve as a guideline for the development of tests and the writing of manuals.

- **COTAN does NOT provide quality marks and/or advice which tests to use and which not**
  - Responsibility of the psychologist/test user!

Trends and developments (1)

Computer-based and Internet-delivered testing

- **In 2009 Dutch Rating System for Test Quality was revised, this trend was among other things the reason for revision**
  - the texts of most criteria were adapted to apply to both paper-and-pencil tests and computer-based tests
    - specific questions with regard to computer-based testing were added
    - within existing questions, more specific descriptions of required details with regard to computer-based testing were added
  - Especially, criteria 2 *Quality of the test materials* needed revision

Trends and developments (2)

2. Quality of the test materials

- Specific questions are added (different items for P&P tests and CB tests)
  - standardization of the test:
    - for adaptive tests, decision rules for starting the test, selecting the next item, and ending the test must be specified
  - scoring; computerized or an objective scoring system?:
    - information should be provided to check correctness of scoring
    - with CBT, this information is usually not described in the manual
  - software design:
    - errors caused by improper use can be avoided?
    - precautions like switch off hotkeys, no access to hard drive, accessing other software
    - Internet tests: manual should contain a description of precautions the test taker should take

Trends and developments (3)

2. Quality of the test materials

- Specific questions are added (different items for P&P tests and CB tests)
  - instructions for test taker:
    - to avoid ‘errors’ because the test taker doesn’t know how software works
    - e.g., available time per item, explanation on adaptive testing
  - quality of the design of the user interface:
    - e.g., consistent and well-organized layout, readability of the information on the screen and the use of colors
  - test security:
    - test developers should do as much as possible to secure access to the test, test content and test results

Trends and developments (4)

Computer-based and Internet-delivered testing

- **Current project: *Translation and adaptation of the ITC Guidelines on Computer-based and Internet-delivered Testing***
  - Possibly: *General Standardised Guideline for Computer-based Testing*
  - Aim: to increase readability for test users
    - e.g., by adding clarifications and examples

Example:

1.a.1.2: "*conduct adequate usability testing of the system requirements using the appropriate delivery platforms to ensure consistency of appearance and delivery.*"

-Addition:

Also consider other platforms than desktop computers and laptops, like tablets.

Trends and developments (4)

Computer-based and Internet-delivered testing

- **Collecting data is a big challenge for practice**
  - and time- and money-consuming!
- **However, data collection via the internet might offer a solution**
- **>> Unproctored vs proctored data collection**

Question from test publishers:
Is it ok when we collect data unproctored?

Trends and developments (5)

Computer-based and Internet-delivered testing

Question from test publishers:
Is it ok when we collect data unproctored?

- **Personality/attitude vs intelligence**
- **Test administration: unproctored or proctored?**
- **Data collection sample representative for population of interest?**

- **For research, more emphasis on equivalence P&P tests and CB tests, and unproctored and proctored test administration.**

Trends and developments (5)

Short(er) tests and questionnaires

- **number of items vs bandwidth of the construct**
  - Example1: broad construct, 5 items, alpha = .90 (items were similar)
    - Is it ok?!
  - Example2: small construct (like intention to leave), 3 items, alpha = .83
    - How many times can you ask the same question over and over again?

Trends and developments (5)

Continuous norming

- **Rating system wrt classical norming:**

	important	less important
Good	N ≥ 400	N ≥ 300
Sufficient	300 ≤ N < 400	200 ≤ N < 300
insufficient	N < 300	N < 200

- **Continuous norming procedure uses the information from all available groups to construct the norms for a specific group, which results in more accurate norms than classical norms**
  - Continuous norming is seen as a method to decrease sample size(s)

Trends and developments (5)

Continuous norming

- **Rating system wrt continuous norming:**

	important	**less important**
Good	N ≥ 150	N ≥ 100
Sufficient	100 ≤ N < 150	70 ≤ N < 100
insufficient	N < 100	N < 70

- NB1: guidelines for sample size of subgroups when 8 subgroups are used
- NB2: only attention for standard error of the mean, not for other moments of the distribution
- NB3: assumption that statistical assumptions are true

Trends and developments (5)

Continuous norming

- **Problem:**
  - Only one specific example, there are too many different models
    - Different number of subgroups and different assumptions
- **Therefore, rating system >> test author should prove equivalence of used sample size with sample size in classical norming**
  - Test authors need more concrete guidelines on how to do this
- **Aim project: to provide more concrete guidelines**
  - it's a challenge, since a lot is unknown and there is limited experience
  - >> research opportunities

Trends and developments (5)

Fairness

- **From a social perspective more emphasis on fairness**
  - Especially in testing
- **Current rating system already has some fairness related items**
- **Aim project: to provide one clear overview wrt conducted fairness research with the instrument that is being rated**
- **For research, more emphasis on DIF analyses and techniques**
  - More emphasis on effect size indices