TST Validation Overview

Tests for Selection and Training a commentary
on the history and validation

THOMAS INTERNATIONAL

TST results from more than 20 years of research and development in computer based psychometric testing, carried out on behalf of central government agencies by the Human Assessment Laboratory of the University of Plymouth.

During that time, members of the laboratory published over seventy reports and papers, of which 36 were devoted particularly to the construction and validation of a test series that began as The British Army Recruit Battery (BARB), developed into The Navy Personnel Series (NPS), and graduated to The Army Regular Commissions Series (ARCOM).

Item Generation Theory

TST comprises 5 tests, each of which demands that the subject complete one type of task in each test for a relatively short period of time. This approach ensures that a number of trials of the same class of task, whose individual items differ in difficulty only within a certain range, can be used during test construction.

The tests consist of computer generated items from a programme that constrained their difficulties. These difficulties were predicted from characteristics of the items observed in performance models. Each test contains a number of trials of one type of item.

From knowledge that some items can be made to be more difficult than others, the series was at first constructed at two levels, initial and advanced entry level.

- Initial entry level means that the tests are suited to school leavers who are entering the employment market for the first time, or retraining, having not first taken a college or university qualification.
- Advanced entry level means that the items were constructed at a higher level.

After successful trials with these two versions, a single, mid-range version was devised to cover the whole spectrum from standard school-leavers to executive level candidates.

The initial test battery underwent standardisation trials for two years before being released. This was called the Combined Version of the Navy Personnel Series (NPL) Tests and this combined version was the prototype for the TST Series.

BASIC PRINCIPLES OF ITEM GENERATION THEORY

- What makes a mental task difficult?
- Isolate the origin of difficulty
- Derive performance models
- Translate models into test construction
- Manipulate the rules
- Set test items at one level of difficulty
- Compile different subtests

Fluid Intelligence

"Fluid intelligence is involved in tests that have very little cultural content, whereas crystallised intelligence loads abilities that have obviously been acquired, such as verbal and numerical ability, mechanical aptitude, social skills, and so on." CATTELL (1983)

Because ability tests are designed to focus on constructs that are not tangible, but rather hypothetical, they cannot be directly measured. Therefore, an individual's level of success in dealing with particular problems is used to infer a degree of general ability in that area.

Intelligence is seen as one of the most important factors in distinguishing between individuals' levels of ability and their potential.

Before the era of tests and measurements intelligence meant "the ability to profit from experience", implying the ability to behave adaptively, to function successfully within particular environments.

"Intelligence" has proved to be notoriously difficult to quantify and attempts to independently define the term have become intertwined with the techniques developed for its measurement.

The dominant use of intelligence tests was as a predictor of academic success, from which we get the definition Intelligence Quotient (IQ). This is an age related measure of intelligence level, where a person's mental age (as determined by a standardised test) is measured against their chronological age.

Modern intelligence tests attempt to measure intelligence as a general ability factor -"g" which can be further sub divided into fluid ability - "gf" and crystallised ability "gc".

Fluid intelligence is pure intellectual speed and power, reflecting the efficiency of the flow of information through the brain and is assessed by the ability to solve novel problems creatively.

Crystallised intelligence relates to learnt factors, and is assessed by tests based on facts and the ability to utilise facts.

"Possibly one of the soundest models to come from out of the structural approach is that proposed by Catell in 1971." (TAYLOR, 1997)

Test Content

The content of the TST Series has been dictated by three principles:

1. The first is that all tests have to be clearly defined in the literature by published work that reveal those aspects of human cognitive performance captured by the test items. In short, theory has to prescribe in practice what cognitive qualities the items demand of people; and what makes these demands progressively more difficult to fulfil.

2. The second principle in the construction is that these tests have a long history as types of cognitive tests which measure known cognitive abilities used in work, training and educational contexts. They were therefore designed to demand from applicants the following essential cognitive qualities:

- Constant attention and concentration
- Memory for task procedures
- Accuracy of decision-making
- Speed of processing information

3. The third principle employed in the development of the TST is that there should be a minimal amount of knowledge / educational level required to take the tests. In order to complete the test series competently, test-takers only need to know the order of letters in the alphabet; to recognise letters in both upper and lower case; to understand simple comparative adjectives such as heavier and shorter, to use negatives to change the meaning of simple sentences, to count up to 30 and to subtract two numbers not greater than 30.

The acquired knowledge demands of TST are, therefore, no more than functional literacy and numeracy. As a result, TST may be among the very few tests ever designed to satisfy the more obvious needs of employers who pursue equal opportunity policies in selection and training.

They may also be described as minority conscious; that means they are constructed in such a way as to minimise the effects of educational disadvantages that are almost invariably shown to be associated with achievement test performance and minority group status.

TEST SPECIFICATION & THEORETICAL BASIS

A literature search was conducted to survey work in the areas of cognitive psychology. The resulting list of references was reduced to three databases of over 300 studies from which information was obtained to formulate a theoretical basis for the task specifications needed.

THE DETTERMAN BATTERY

Douglas Detterman viewed intelligence as a complex system of a finite number of independent variables. He suggested that if it were possible to measure each of these variables separately then the combination of these measures should be predictive of more complex tasks. This was the theoretical underpinning of the test system.

TASK SPECIFICATIONS

The content of the test battery was chosen from across the spectrum of cognitive abilities to be representative of the kinds of processes believed to be the building blocks of higher order abilities and "intelligence" and was designed to provide an index of "trainability".

THE LAMP PROJECT

The working definition of "trainability" owed much to the work carried out by the Learning Abilities Measurement Program.

The goal was to understand the components of the human information processing system that enable learning to occur.

The model they adopted proposed that individual differences in the ability to learn resulted from four sources:

- Declarative (factual) knowledge – "what"
- Procedural (strategic) knowledge – "how"
- Speed of elementary information processing
- Attentional working memory capacity

FIRST STEPS TO THEORY VALIDATION

It was predicted that tests could be manufactured from algorithms whenever these algorithms determine a theory-driven item-production system. The production system enables a number of measures of information processing, whose common ground is allocation of attentional resources to solving problems in short-term and working memory. In this context, they will correlate moderately well with each other, with tests of declarative knowledge, and with procedural learning tasks. They would be expected to correlate with other tests also emphasising working memory and to predict differentially those criterion tasks with which they had content overlap.

Apart from construct validity, the theory is testable more immediately from internal test characteristics. It follows that if the algorithms produce the test items randomly within a given block size, and that there is time for completion of at least one block, parallel forms of the tests should be produced consistently, with only random variation in descriptive test characteristics from form to form.

When the algorithms are thought to function properly, producing stable internal features and confirmed construct validation, the final step is to test the structural models of the tests with as many subjects as possible, using appropriate linear methods.

The quickest and least expensive trial of internal test features proved to be the computer generation of parallel forms of paper and pencil analogues. A number of these were produced.

From 1989-1993, many studies were completed on the feasibility of use, reliability and validity of TST. Several sets of tests were used for validation including a paper and pencil analogue of the British Army Recruit Battery at basic literacy level and the advanced form of TST. A final combined form was constructed using most of the advanced items and the Working Memory test in place of the more difficult Alphabet Forward and Backward test.

During this period procedures for test administration and scoring were developed and the results of the trials with TST were related to training performance in a wide variety of jobs spanning the whole of the occupational spectrum.

Reliability

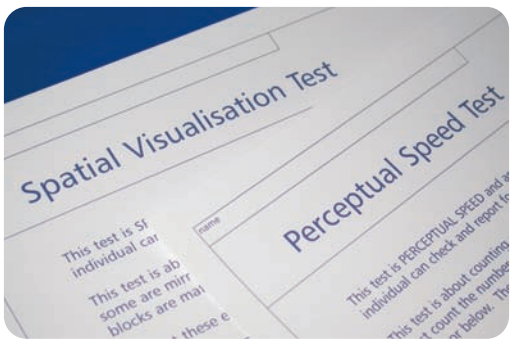
Reliability is an index of how well an instrument (in this case the TST) is measuring the same thing every time it is used. It is therefore a measure of consistency and a reliable test will produce the same result for an individual (or the same pattern of results for a population) each time and, within the test itself, each test item will produce consistent measurements on each administration of the test.

There are two main ways of measuring reliability:

- By comparing sets of items selected from the test at one testing (e.g. comparing one half of the item responses, chosen randomly, with the other half). These results give an indication of the internal consistency of the test.
- By testing the same subjects twice with a time interval between testing and comparing the scores obtained from the initial test and the re-test. This measure is known as test-retest reliability. Generally speaking, the greater the time interval, the lower the index of agreement.

The results of comprehensive reliability trials show that the TST series has very good internal consistency on a single trial and satisfactory test-retest reliability. When new versions of the tests are generated there is also very high consistency between versions ("parallelism"), which means that new versions need not be subjected to intensive re-standardisation and norm-adjustment exercises before they can be used.

The TST can also provide a General Trainability Quotient (GTQ), indicating how well an individual is likely to perform in training contexts, by the calculation of a composite score across all of the tests. When this is done, the reliability of the composite has been shown by Wright to be not less than .95. (Wright, D.E. 1992. IRT modelling using latent variable generalised linear models. HAL Technical Report 3-1992 [APRE]. Human Assessment Laboratory, University of Plymouth, Devon, UK).



Validity

CONSTRUCT VALIDITY

Construct validity is an index of how well an instrument is measuring what it is supposed to be measuring. It is obtained by examining the structure of correlations among test scores. Principle components factor analyses were used in the validation studies for TST and related tests, including one independently carried out by Weldon (1993) on 886 Army trainees tested with a computer-based version.

The results of analyses of the construct validity of tests in TST most commonly reveal a single factor with pronounced loadings on the three tests associated with working memory: Reasoning, alphabet manipulation (Working Memory test) and Number Speed and Accuracy - with some specificity shown from the Spatial Visualisation and Perceptual Speed tests.

A total-score GTQ provides a measure of general, mainly fluid, intelligence.

CONCURRENT VALIDITY

Concurrent validities are correlations of tests that are not the same tests as those in the TST Series, but are other cognate tests, which are administered at or about the same time.

The TST tests were correlated with various Royal Navy (RN) entrance tests given at the same time. The concurrent results show that they have moderate inter-correlations with RN test composites. This is not considered a disadvantage, because they are thereby able to add to predictive validities already found with Navy tests. The results show some overlap, but not enough to render either the TST or Navy tests redundant when used in combination

CONVERGENT AND DISCRIMINANT VALIDITY

The meaning of test validity can be extended, by asking if new tests correlate with other tests that are supposed to measure the same attributes (convergent) and fail to correlate with those that measure other qualities (discriminant).

In the TST validation studies, results showed that in almost every case verbal, mathematical and spatial tests correlate most highly with other tests in the same domains, and show least relationships with tests measuring very different types of performance.

Another aspect of concurrent validity is demonstrated whenever a pattern of scores for a group serves to confirm differences in performance levels between groups. Results showed well-defined differences in performance among different tests, depending on the branch of the trainee. Because these personnel were allocated to these branches without reference to the TST tests, the series itself shows that it is responsive to the predefined differences in aptitude and interests of the various groups. There are clear differences in test profile means among clerical, technical and service occupations.

PREDICTIVE VALIDITY

TST data was correlated with global stage criteria to evaluate test efficiency in allocating personnel to occupations. Because of the high numbers of subjects that have been followed up, the correlations were determined to be stable estimates of efficiency, provided that training syllabuses and trainee management procedures were followed. The correlations were all high when corrected for populations of applicants and were significant at $p<0.01$.

INCREMENTAL VALIDITY

New tests have to show that they can reach the same level of efficiency at less cost, or add significantly to the predictive power of the tests already in use to demonstrate incremental validity. The TST tests are unique in predicting practical knowledge acquisition in every-day learning on the job and the low-cost maintenance and parallelism are advantageous.

Task Types

Item-generation algorithms were constructed for each of the 5 tests, which cover four of Carroll's second-order psychometric constructs. All contribute to a third-order general intelligence as defined in Carroll's (1986a,b, 1993) structure of intellect model derived from psychometric test inter-correlations.

In brief, the tests consist of generic types. Each test is illustrated and referenced by its psychometric factor, knowledge requirements and by its "cognitive model". These reference frames enable the reader to refer to the technical literature.

PERCEPTUAL SPEED

Testing Functions: The capacity to recognise details in the environment, incorporating the perception of inaccuracies in written material, numbers and diagrams, the ability to ignore irrelevant information and identify similarities and differences in visual configurations. This test assesses how quickly and accurately an individual can check and report for error/accuracy. It is a task of semantic encoding and perception. A high score would suggest the ability to mentally match the features of letters and the meaning of symbols. It would also indicate the ability to detect misfits.

Knowledge: Alphabet letters in upper and lower case.

Psychometric factor: General Speed Gs.

Performance Model: Semantic encoding and comparison. (Sternberg, 1966; Posner, Boies, Eichelman & Taylor, 1969; Hunt, Lunneborg & Lewis, 1975; Irvine & Reuning, 1981; Irvine, Schoeman & Prinsloo, 1988).

REASONING

Testing Functions: The ability to make inferences, to reason from information provided and to draw correct conclusions. This test assesses the ability of an individual to hold information in his/her short-term memory and solve problems after receiving either verbal or written instructions. A high score would suggest fluent verbal reasoning skills.

Knowledge: Comprehension of simple sentences; use of comparatives and negatives in assigning meaning.

Psychometric Factor: Fluid General Intelligence Gf.

Performance Model: Decisions based on structural determinants of sentences (Clark, 1969, Clark & Chase, 1972; Evans, 1982).

NUMBER, SPEED & ACCURACY

Testing Functions: This is a test of numerical manipulation and a measure of basic numerical reasoning ability. It can therefore be used as an indicator of the degree to which an individual can work comfortably with quantitative concepts. It assesses the ability to work in environments where basic numeracy is required and wherever attention and concentration are required regarding numerical applications.

Knowledge: Order of numbers to specified range. Number facts for one and two-digit subtraction pairs within the range specified.

Psychometric Factor: General Memory Capacity (with Numerical Specific) Gm.

Performance Model: Decisions based on number retrieval (Moyer & Landauer, 1967; Groen & Parkman, 1972; Parkman, 1972).

WORKING MEMORY

Testing Functions: The ability to hold information that has been previously processed, while simultaneously processing and assimilating incoming information. It is a test that makes demands on reconstructive memory process using the letters of the alphabet. It is central to many everyday tasks such as reading, making sense of spoken discourse, problem solving and mental arithmetic, as demonstrated by reliable correlations between tests of working memory and a range of real world skills.

Knowledge: Alphabet letters in sequence from first to last.

Psychometric Factor: General Memory Capacity Gm.

Performance Model: Reconstructive memory- task. (Hockey, Maclean & Hamilton, 1981; Hockey & Maclean, 1986; Woltz, 1987).

SPATIAL VISUALISATION

Testing Functions: The ability to create and manipulate mental images of objects. This test correlates well with tests of mechanical reasoning and assesses an individual's ability to use mental visualisation skills to compare shapes. It relates to the ability to work in environments where visualisation skills are prerequisites for understanding and executing tasks. It assesses the suitability of an individual for tasks such as design work, where the individual must visualise how shapes and patterns fit together to form a whole.

Knowledge: Recognition of shape and its mirror image.

Psychometric Factor: General Visualisation Gv

Performance Model: Spatial rotation of two-dimensional symbols. (Shepard & Metzler, 1971; Just & Carpenter, 1985; Bejar, 1986a,b.c).

TST ENTRY LEVEL CRITERIA

The Candidate has to be able to:

- Know the alphabet and the order of its letters
- Understand the difference between upper case and lower case letters
- Compare simple adjectives like heavier and lighter
- Count up to 30
- Subtract two numbers not greater than 30
- Recognise a shape and its mirror image

Perceptual Speed

Measures ability to:

- Recognise details
- Identify similarities
- Check for error and accuracy
- Ignore irrelevant information
- Detect misfits

Spatial Visualisation

Measures ability to:

- Manipulate mental images of objects
- Compare shapes
- Visualise how shapes and patterns fit together

GTQ

Measures:

- Fluid intelligence
- Concentration
- Response to training
- Mental processing speed
- Fast track potential

Reasoning

Measures ability to:

- Make inferences
- Reason from information provided
- Draw correct conclusions
- Hold information in short term memory and solve problems

Working Memory

Measures ability to:

- Hold previously processed information, whilst assimilating incoming data
- Retrieve information
- Make sense of information

Number Speed and Accuracy

Measures ability to:

- Manipulate numbers
- Concentrate on numbers
- Work with quantitative concepts

Summary

- TST is an independently validated battery of normative ability tests demonstrating all forms of validity.
- TST measure fluid intelligence, as opposed to accumulated knowledge or skills, using a battery of five sub-tests.
- TST deliver a general training quotient (as a measure of trainability) and a report on the individual's areas of development potential.
- TST provide a reliable, accurate and valid means of identifying how quickly a person can learn and retain new skills and procedures.
- TST is objective, fair and discriminates positively.
- TST has a low entry level and is applicable at all levels of the organisation.

www.thomasinternational.net

Thomas International
Harris House
17 West Street
Marlow Bucks SL7 2LS

Telephone 01628 475366
Facsimile 01628 524226
Email info@thomas.co.uk

Ref: TSTVAL04.06 V1