Testing and Assessment Practices: Journey and Conceptual Reflections

Ajoy Konar [1]
Ayesha Martin [2]
Sombala Ningthoujam [3]

[1] Institute of Banking and Personnel Selection, Mumbai
ajoykonar10@gmail.com

[2] Institute of Banking and Personnel Selection, Mumbai
ayesha.martin@ibps,in

[3] Institute of Banking and Personnel Selection, Mumbai
Sumbala.ningthoujam@ibps.in

Abstract

Test instruments can be measures of knowledge, skill, aptitudes, vocational interests or general intelligence. The data generated by these instruments are used for decision making in varied contexts from diagnosis to selection with consequences for individuals, groups, organisations and the public at large. Hence application of scientific and professional best practices to the selection and evaluation of tests used, test administration, scoring and interpretation to counter the effects on mental health as well as threats or bias in the reliability, validity and fairness of assessment, is important. Appropriateness of test administration procedures and calibration of tests ameliorates performance especially in high stakes test settings. Test design (e.g., timing and pacing, test length) and test administration (competence of test proctors, effectiveness of processing formalities at test centres) should be sensitive to and commensurate with the growth in the assessment field. The aim of the article is to inform understanding of psychological testing in the field of assessment for test developers, educators responsible for instruction and assessment, employers and those impacted by decisions taken on the basis of psychological tests and policymakers and bodies who make decisions with regard to psychological assessment and evaluation.

Keywords: High-stakes Assessment, Technology, Test Anxiety, Standardised Tests, Evaluation

Introduction

The concept and practice of testing has evolved over time. Assessment, which began mainly for the purpose of identifying individual differences in the areas of intelligence, aptitude, personality and educational achievement owing to the need felt in the education sector, has grown. From a very early stage, teachers have assessed their students based on their performance during the teaching process or by administering a process of evaluation at the end of the teaching programme. The use of valid testing and assessment methods for selection

gradually extended to industries and business houses and corporates in public and private sector to identify the right talent for selection for/ placement in various jobs.

Assessment takes many forms and can be applied in a range of clinical, educational, organisational and forensic settings. It stems from practical needs which can be broadly classified into *research, diagnosis, development, prognostication (prediction) and selection*. Tests provide a source of information for assessment and are of varied types. These include inventories, checklists, projective techniques etc. to measure interests, achievement, aptitude, intelligence, traits etc.

The assessment of individuals to measure traits, abilities, achievements and aptitude has been a challenging task since early 19th century (Freeman, 1962). The objectives of the assessment have to be clearly kept in view at all times for relevant psychometric decisions as well as planning operational strategies (Deshpande, 1986). Methodology, item types, delivery systems have to be designed and analysed with a view to ensuring sufficient reliability, validity and fairness in assessment systems.

While assessments serve many purposes, these can also be a source of anxiety for test takers, either motivating performance for some or hampering performance for others. Along with individual differences of trait or state anxiety (Derakshan & Eysenck, 2009), certain factors such as age (Byrne, Davenport, & Mazanov 2007; Klinger et al. 2015), gender (Putwain & Daly, 2014, von der Embse et al., 2018), academic ability (Hembree, 1988; von der Embse et al., 2018), having special educational needs and disability and the use of access arrangements (Heiman & Precel, 2003; Nelson & Hardwood, 2010) as well as perceived importance of the examination i.e., 'examination stakes', perceptions of familiarity, difficulty of the test assessment (Bonaccio & Reeve, 2010; Cizek, 2001; Hembree, 1988; Pekrun et al., 2004) can predispose individuals to experience test anxiety before and during. While individual differences do impact experience of test anxiety, assessment-specific design features such as item types and item formats (e.g., multiple choice versus constructed response items) too may induce test anxiety. While optimal measurement of the construct (skill, behaviour or knowledge) that

is intended to be measured should not be sacrificed, the focus of test developers should be to design assessments in a manner to be less anxiety-eliciting for test takers so as to gain a 'true' picture of ability.

Large scale group tests provide a wealth of data for assessment- test data (ability, trait, achievement, aptitude) and demographic data and thus have tremendous implications for both policymakers and stakeholders. The aim of this paper is to provide through a discussion of historical perspectives and developments in psychological testing, a practical account of the evolution and application of core concepts and various testing approaches and methodologies. It aims to link (principles) 'what needs to be done' and (practice) 'how it has been done' and 'could be done' reiterating the continuing relevance of core principles to the process of large-scale psychometric testing and assessment in diverse multidisciplinary settings.

Methods of Assessment and Testing: Classification of Tests

Psychological testing must be understood as a dynamic process. Though measures of personality and intelligence are the most prominently easily recognisable psychological tests, there are, in fact, many types of psychological tests for varied purposes e.g., college examinations, vocational inventories, aptitude tests etc.

A test could comprise a single scale or multiple scales. For example, in the case of aptitude testing a single test or multiple scales in the form of a test battery may be used depending on the spectrum of abilities/ traits to be assessed. The achievement test has been prevalent all over the world for admission to educational institutions and universities as well as for the award of Degrees, Certifications and Diplomas, the implicit assumption being that the educational institutions have taught the subject matter directly. Screening and diagnostic measures have also been typically used in the area of mental health for assessing mental health status and needs. In organisational settings, tests have been used for selection, training and development of employees at various levels. Thus,

psychological tests are used by different stakeholders for different purposes in different settings. The most common classification of tests is:

1. *Intelligence Tests*: Instruments designed to estimate an individual's general intellectual level.

2. *Ability / Aptitude Tests*: Instruments/ scales to measure the capability for a relatively specific task or type of skill e.g., to predict success in an occupation, training course or educational endeavour.

3. *Achievement Tests*: Instruments to determine a person's degree of learning or success in a prescribed subject matter i.e., to determine how much of the material has been absorbed and mastered.

4. *Personality/ Affective Measures*: Instruments for measuring traits, qualities or behaviours that determine individuality. Checklists, inventories and projective techniques such as sentence completion and inkblots.

5. *Interest Inventories*: Instruments for measuring an individual's preference for certain activities or topics and thereby often used for determining an occupational choice

6. *Psychomotor Ability or Performance Tests*: Instruments of ability requiring primarily motor, rather than verbal, responses, such as a test requiring manipulation of different objects or completion of a task that involves physical movement

A number of factors determine the type and complexity of test developed and used. For instance, there is an important connection between the objective of testing on one hand and type of test on the other. In selection, group tests can be used for prognostication (prediction) e.g., in rank order placement (to assess an individual's performance against other individuals in the group) or to assess which students will be promoted to the next year. Individual tests could be used for evaluation to indicate the scholastic progress of the child to the teacher

or for diagnostic purposes to determine if a student requires special teaching in mathematics. An understanding of classification of tests thus provides a framework, which is useful in guiding test developers in their approach to test selection and development.

Chronology of Test Development: Historical Perspective

Various developments in the 19th century and early 20th century laid the foundation for testing as we know it today. These include the creation of laboratory tests of sensory discrimination, motor skills and reaction time by experimental psychologists in Germany as well as measurement instruments and statistical techniques generated for the study of individual differences by Galton and his students, (Urbina, 2014). James Cattell in 1890 published a classic paper 'Mental tests and Measurements' which is widely regarded as laying the foundation of modern testing. Among those who made monumental contributions to the field were Clark Wissler who sought to demonstrate that test results could predict academic performance and Strong who devised the Vocational Interest Blank (since revised and still widely used). Failures in some of these experiments resulted in a widespread perception that Galton had been wrong to infer complex abilities from simple ones (Gregory, 2015).

The void left after the abandonment of the Galton tradition was filled by what is popularly perceived as the first modern test of intelligence. Developed by Alfred Binet and Theodore Simon in 1905 it can be termed as a major breakthrough in psychological testing. The Binet - Simon scale as it was called was designed to identify Parisian school children aged 3-13 years, who needed special educational attention (Binet & Simon, 1905). Their aim was classification not measurement. There were subsequent revisions of the scale in 1908 and 1911. In 1908 a revised scale consisting many new tests was added and the major innovation was the introduction of the concept of mental level. The Binet-Simon scales were standardised tests which helped solve a practical social need of identifying those who needed special schooling.

In 1916, Terman and his associates at Stanford University published revised scales of the Binet-Simon test, called Stanford–Binet intelligence scale. The Stanford-

Binet scale has remained the standard of intelligence testing for decades and the latest version was completed in 2003. The Weschler scales (1949 and 1955) which provide in addition to a total IQ score, subtest scores such as verbal and performance IQ have been a popular alternative to the Stanford-Binet test of intelligence (which provides only a global IQ score). Of all standardised tests the Weschler Adult Intelligence Scale (WAIS) has received the most attention with regard to cultural adaptation (Puente et al 2000). Special assessment techniques are essential when working with test-takers from different cultures. Language is one such cultural variable that can significantly impact performance on a test. In fact, translation of the Binet-Simon scale with minor changes was done by Henry Goddard to make it applicable to American children. Despite the best intentions, some applications of psychological tests which took place could be deemed controversial as the humanistic idea with which they were developed was lost sight of.

Intelligence scales were then employed for the quick and accurate classification of army recruits. In 1917, the United States entered the World War. For the purpose of classification and alignment Yerkes suggested that all 1.75 million Army recruits should undergo intelligence tests. This was accepted and he assembled a committee to develop group tests for assessment of Army recruits. Yerkes, Goodard and Terman along with other members of the committee developed the well-known Army Alpha Test (based on the unpublished work of Otis) and Army Beta Test (non-verbal test). The format and content of these tests influenced intelligence testing for years. These tests provided psychologists with tremendous test data and experience in psychometrics making assessment a more scientific practice (Gregory, 2014). Group tests which were slow to catch on being laborious to score, picked up rapidly in the US as the Army Alpha and Army Beta tests ushered in a new era for group tests in industry, schools and colleges.

Thus, the work of Binet gave impetus to test makers, psychologists, scientists across the globe (Sandra and Leslie, 2007) and psychological testing grew in popularity, stature and practice. The newfound popularity of paper-pencil tests of intelligence for groups in turn influenced the growth of testing for scholastic aptitude and college entrance examinations. In fact, the Scholastic Aptitude Test

(SAT), Graduate Record Examination (GRE) could also be traced back to these early efforts of Yerkes, Otis in mass scale army testing.

The development of aptitude tests however lagged behind till World War II when there was a pressing need to select candidates for specialized roles. The armed forces developed a specialised aptitude battery which proved invaluable for selection of pilots, navigators and others as there were lower washout rates (Goslin, 1963). Aptitude tests are useful not only in job selection but in schools and colleges as well for helping students identify courses which best 'fit' them. Many aptitude tests have arisen out of Thurstone's tests of seven primary mental abilities. One of the most well-known of which is the General Aptitude Test Battery (GATB). Devising such tests demand a tremendous investment of time and effort and creating such tests for every job is not practical. Hence 'jobs' have been assembled into 'job families''.

Personality testing too gained impetus with armed services testing. During World War I, the U.S. military wanted a test to help identify soldiers who would not be able to handle the stress associated with combat. To meet this need, the American Psychological Association (APA) commissioned an American psychologist, Robert Woodworth, to design such a test, which came to be known as the Personal Data Sheet (PDS). Woodworth subsequently developed an instrument to detect Army recruits who were susceptible to psychoneurosis. Personality testing began with assessment methods such as free association. During the 1930s, interest also grew in measuring personality by exploring the unconscious. With this interest came the development of two important projective tests: the Rorschach Inkblot Test (for abnormal subjects) and the Thematic Apperception Test (TAT). Gradually the modern objective and empirical approach to personality testing proliferated. Various schedules, scales, self- report inventories such as the Minnesota Multiphasic Personality Inventory (MMPI) emerged as vehicles for studying personality.

Chronology of Significant Trends in Large Scale Standardised Test Development

Milestones in large scale testing transpired in the US and UK in the 20th century. In the US, it was the establishment of a centralised testing body for college admission. This was done to rationalise the admission processes and bring higher quality to testing in education, by addressing the dual concerns of subjectivity in scoring and variation in difficulty from one administration to the next which characterised the existing process (Hubin 1988). Two broad streams of research were undertaken i.e., to supplement (not replace) subjective essay tests either through the use of achievement tests or aptitude tests (Hubin 1988). Investigative efforts centred on the Army Alpha Test and led to the derivation of the large-scale standardised tests for undergraduate admission and graduate level studies such as the Scholastic Aptitude Test (SAT) and Graduate Record Examination (GRE) test.

Selection for civil services in the UK also underwent a significant change during World War II. Selection had traditionally been made on the basis of subject knowledge examinations (Method I). As education was disrupted during the war, this method was temporarily replaced with the administration of aptitude tests (Method II). Eventually Method II supplanted Method I (Deshpande, 1988).

Much of the developmental work in the scientific methods of selection can be traced to the efforts of early industrial psychologists to support the military through two World Wars (Scroggins et al, 2008). The success of psychological tests in finding and predicting merit was well documented by military psychologists in the United States and other countries by the 1940s. The need to classify and select large numbers of recruits for military service led, in 1940, to the formation of the Committee on Classification of Military Personnel. The development and dissemination of the Army General Classification Test to replace the U. S. Army's system of alpha and beta developed in World War I was a major development in personnel selection and classification testing. Psychologists developed aptitude tests and tests of special skills as well as assessment centre techniques, and

set the stage for the later development of the Armed Forces Qualification Test (AFQT) and the Armed Services Vocational Aptitude Battery (ASVAB).

The Armed Services Vocational Aptitude Battery (ASVAB) was introduced in the 1970s by the US Department of Defense (DoD) for selection of US Military personnel. The rigour and validity of the ASVAB has been such that the cognitive level of personnel across the years could be compared and trends in military needs and civilian supply could be anticipated (Maier, 1993). It became the nation's largest personnel system—processing over 800,000 recruits annually- using psychological tests for justifying selection, placement, and training decisions, a practice which became institutionalised and accepted. Over time, substantial changes in usage, mode of delivery (Sands, Waters, & McBride, 1997; Segall & Moreno, 1999), and in the psychometric theory behind the construction and scoring of the ASVAB have served to enhance its usefulness.

One of the most highly regarded standardised test batteries is the United States Training and Employment Service General Aptitude Test Battery (GATB), first published in 1947. The GATB consists of 12 tests which measure nine aptitudes- cognitive, perceptual, and psychomotor skills thought relevant to the prediction of job performance. The battery is used for vocational counselling and occupational screening. In 1983, Prof. John Hunter's research report on Test Validation for 12,000 jobs: An application of job classification and Validity Generalisation Analysis to the General Aptitude Test Battery (GATB) was published by the Division of Counselling and Test Development, U.S. Department of Labour. It exploded the myth that the GATB was valid for some jobs and not others. The GATB was found to significantly predict job performance for all the job titles. Efforts to validate the GATB have also been undertaken in India and one such effort was by Dolke (1978) who adapted seven paper-pencil tests to the Indian setting. Over the years, hundreds of studies have been undertaken in order to assess the validity of the GATB and its nine scales. The GATB has been found to be a moderately valid predictor of job performance and is widely used and well regarded even today.

The Differential Aptitude Test (DAT) is a multiple aptitude battery. Designed to measure junior and senior high school students' and adults' ability to learn

or succeed in certain areas, the test published in 1947 is suitable for group administration and is primarily for use in educational and vocational counselling but can also be used in employee selection. Correlation of the DAT with several well-known tests for aptitude such as ACT and ASVAB as well as with achievement (CAT, CTBS, etc.) and with cumulative GPA revealed a positive relationship between the DAT and achievement tests. The DAT was also found to serve as a good predictor of GPAs. The DAT has remained one of the most frequently used batteries owing to its efforts for improvement of psychometric quality.

In March 1953, a conference at which 12 graduate schools of business were represented agreed that a nationwide testing program would be useful. The test was called the Admission Test for Graduate Study in Business until 1976. In that year the test name was changed to Graduate Management Admission Test (GMAT). The GMAT was first administered in February 1954 to about 1,300 prospective students of graduate schools of business. In order to ensure that the test scores may be relied upon by graduate management programme admissions officers to supplement other data about applicants, such as previous academic performance, the following conditions were set namely that the test must (i) measure abilities that are relevant to successful performance in graduate management school and that are developed by a wide range of educational experiences, (ii) be sufficiently long to provide a reasonably dependable measure, (iii) be administered under uniform, secure conditions, and (iv) be scored accurately and reported promptly in a convenient form. In recent years, GMAT examinees were asked to answer biographical data questions and their answers were transmitted as part of the GMAT report to schools that they designate. A study of the representation of major undergraduate fields in the GMAT examinee group revealed that graduate students in management are drawn from a wide spectrum of undergraduate major fields (Shrader, 1984).

The tests discussed thus far are standardised tests, designed as 'foot rules' to give a standardised measurement wherever they are used. These have been developed and have evolved thanks to a combination of the following factors in the field of testing (i) theoretical advances (NEO-PI-R: Costa & McRae 1995), (ii) empirical advances (MMPI Butcher, Dahlstrom, Graham, Tellegen & Kaemmer,

1989) and (iii) practical need (SAT: Coyle & Pillow, 2008; GMAT: Oh, Schmidt, Shaffer & Le, 2008).

The summary of these decades of research:

(I) leads to a classification of the purposes for assessing individuals as follows –

(i) Measurement of Ability/ Aptitude for performing a future task

(ii) Measurement of Achievement (attainment of knowledge) imparted during a programme

(iii) Identification of areas for developmental purposes

It is clear that practical applications of assessment will frequently overlap. For example, in the evaluation of a remedial programme for underprivileged children from low-income families, an appropriately designed test for achievement could be used for (ii) and (iii) i.e., classification and development. The data from the test is also likely to serve as a better indicator of scholastic achievement than anecdotal evidence.

(II) reveals a trend in personnel selection away from the achievement type of tests towards aptitude tests to make selection more cost-effective and valid. This is particularly true in cases where learning opportunities are unevenly distributed and these tests also have proven to be fairer across various sectors of society (Deshpande, 1988)

(III) gave impetus to contemporary practices of assessment and testing namely

(i) the design, development and publication of standardizsed objective tests as per specifications by professional bodies such as the American Psychological Association (APA)

(ii) various empirical practices such as systematic job analysis and an alignment of the test with the skills needed, revisions in forms of tests and scoring in keeping with the skills

(iii) changes in mode of delivery

(iv) changes in the psychometric theory behind test construction and scoring e.g., Computer Adaptive Testing (CAT)

Today the field is a vibrant one with the advances in software and technology giving rise to increasingly sophisticated models and practices. Scientific

advances in the field, from the use of Optical Mark Recognition (OMR) technology to Computer Based Testing (CBT), Automated Evaluation Systems (AES), innovations in digital assessments which includes leveraging of algorithms for measurement in terms of both response and process data and advances in statistical tools such as exploratory and confirmatory factor analysis to assess the quality of tests and scores, have and will support the large-scale operational use of testing and facilitate greater precision in test design, administration and scoring and evaluation.

Methods of Large-Scale Assessment

In the context of large-scale assessments, tests are always preferable to other techniques of assessment owing to their demonstrable reliability, validity and objectivity as a form of assessment. Multiple-choice items require examinees to select a response from given options, while Constructed Response items present an item stem and require examinees to construct a response "from scratch." (Lissitz, Hou and Slater, 2012).

With advances in technology and in the field of psychology and psychometrics, operational changes in test construction, administration, scoring and interpretation procedures are introduced from time to time to ensure comparability, objectivity and accuracy. Depending on the needs of the selection system including considerations of cost, reach, security and objectivity, tests both objective (multiple choice format) and subjective (constructed response/ short answer/ descriptive) are administered through different modes as follows –

a. Paper-Pencil format
b. Machine scoreable forms (OMR technology and Onscreen Assessment)
c. Computer Based Tests (CBT)
d. Computer Adaptive Testing (CAT)

Psychological measurement instruments based on the paper-and-pencil conventional test were developed initially for use in World War I to provide a quick and inexpensive method of screening large numbers of recruits (DuBois, 1970). This type of test was designed using procedures of classical test theory (e.g.,

Cronbach, 1990; Gulliksen, 1950), and had roots in procedures that developed around the same time as the paper-and-pencil test (Weiss, 2004).

In order to ensure the efficacy of a test, it is important to consider the goals and objectives of the total selection system (Deshpande, 1986). Alongside the need to match selectees with job requirements and ensure speed in the process of assessment, public accountability and fairness are important considerations. In India, traditionally most assessments and examinations were of descriptive type. However, a shift to objective type testing took place owing to a need for a more robust scientific system to replace the ad hoc one which was in existence in the banking sector. The National Institute of Bank Management (NIBM), an apex level training and research organisation at the time was entrusted with this task. In addition to recommendations for a selection system, it can be credited with popularising objective type testing in India in the 1970s. Subsequently, organisations other than Public Sector Banks (PSBs) adopted objective type testing. Multiple choice items generally demonstrate greater content validity than do Constructed Response items (Newstead & Dennis, 1994). They also show demonstrable objectivity, fairness and impartiality owing to which it was introduced on a mass scale for both employment and admission testing.

Traditionally in objective type paper-pencil tests, candidates indicated their answers by a cross-mark on a separate answer sheet. The responses of these examinations would then be objectively scored using a stencil (scoring sheet/ template). The process was human and time intensive with large volumes of candidates.

To reduce the time spent on the massive support activity of marking and scoring candidate responses, mechanisation of this activity was called for. For objective type tests, after a survey of mechanised scoring systems in countries such as the U.K., Japan and the U.S., Optical Mark Recognition (OMR) or “mark sensing” technology was adopted. OMR technology, a conventional data input system that senses the presence or absence of marks by recognising their darkness on a sheet of paper is an example of a human-computer interaction procedure (Deng et al., 2008). The candidates were required to mark their answers by darkening

ovals in pencil. The darkened ovals were read by the machine as answers and were stored in the computer. The right answer-key was matched with the candidate's responses to generate scores. The answer sheets to be used with the OMR system needed a particular kind of paper, specific ink and precision in printing. This was because the OMR scanner was sensitive to any disruption in the integrity of the answer sheet and this could prevent the answer sheet from being read correctly.

At the time optical scanners were to be imported to India. The use of OMR technology for scoring answer sheets of objective tests in mass level high stakes selection testing was implemented and popularised by IBPS. The system proved largely reliable, consistent and accurate in scoring answer sheets. It eliminated entirely the requirement of data-punching and verification activity carried out earlier. The data were captured and stored in digital format. Software programmes developed for processing reduced the time taken to less than half. It also brought in accuracy to a level of perfection. This technological intervention started in 1985 and continued for three decades. This model reduced the time frame of declaration of results to a great extent. The accuracy of the OMR system was checked by passing the same answer sheets twice add through the scanner after the word and comparing the scores generated for each answer sheets in two passes (Deshpande, 1986).

The ease of scoring also permits score reporting to be accomplished more quickly, thus reducing the lifecycle of a selection exercise and in case of educational testing providing students and teachers with feedback on performance in a timelier manner. It permitted an unprecedented expansion of large-scale testing activities and huge reductions in the cost of testing (Martinez and Bennett, 1992).

Some organisations are still using this technology for large examinations. This technology requires printing of very large number of question papers and also answer sheets on specific paper. Logistical issues remain a big challenge even today.

Objective tests have been in use for decades. It is relatively easier to ensure comparability of scores over time even for tests which are not standardised. This can be accomplished through fixed time, machine scoring of answer sheets, equating etc. (McClellan, 2010). However, the descriptive (constructed response) format of examinations is still prevalent either in part or in full in India. Many academic assessments include a large set of multiple-choice questions to sample students' knowledge in a broad domain and a few constructed-response questions to assess the students' ability to apply that knowledge.

In the traditional format of administering descriptive type tests, a set of questions is made available to candidates and candidates are required to construct (write) their responses to these in a blank answer book. Only a limited number of questions (at the most five to ten questions) can be administered in a single sitting of an examination. Descriptive type examinations thus allow us to ascertain in-depth knowledge on a select spectrum of the domain (very few topics) during the limited test time. Objective type tests however, allow a larger sample of the domain area to be evaluated- larger number of questions can be presented to candidates to respond to in a shorter time period. According to Angoff, (1953); Crocker & Algina, (1986), tests comprised of more items tend to have higher test reliability. Though the tasks that constructed-response questions require of the test taker can be varied and complex skills measured, an individual test taker's performance will tend to vary more from one set of questions to another (Livingston 2009). Thus, while the challenge with objective tests lies in development of a large number of quality MCQs in each domain and ensuring these are relevant to the domain, with constructed response items scoring reliability is the significant challenge.

While descriptive papers are evaluated by human evaluators, objective tests are scored by machines objectively (Attali, Yigal, 2015 & Attali, Yigal, and Sandip Sinharay, 2015) and therefore can be said to be free from human bias. There has been no uniform approach adopted by institutes, schools, colleges and universities with regard to evaluation of descriptive papers. However, efforts are being made by various academic institutes to bring in rigour in manual evaluation of descriptive papers.

There are two basic approaches to the scoring of constructed-response test questions-analytic scoring and holistic scoring. In both cases, the scoring is based on a set of guidelines called a rubric. The rubric tells the scorer what features of the response to focus on and how to decide how many points to award to the response. An analytic scoring rubric lists specific features of the response and specifies the number of points to award for each feature. In the process of holistic scoring the scorer reads the response and makes a single judgment of the quality of the response by assigning a numerical score (Livingston, 2009).

Various practical methods have been adopted to limit or reduce subjectivity. These can range from limiting the number of expert raters to reduce inter-rater difference to adopting the process of moderation of scores through back scoring or double scoring. Statistical analysis of scores on items and the assessment as a whole should also be an integral part of the process. A recent trend is manually scoring physical answer-papers after masking the identity of test-takers from the evaluators.

Traditionally when there is a constructed response (descriptive) component, the result preparation and declaration by testing bodies is impacted as scoring of constructed responses (descriptive answers) takes considerable time. A development in the evaluation of paper-pencil descriptive type tests that has taken place in the last five years, is the process of on-screen evaluation. In this process, the answer-papers of candidates are scanned after masking the student's identity. Scanned copies of candidates' answer-papers are randomly presented to the evaluators in a centralised location for evaluation on a computer system. This process is commonly known as On-screen Marking (Swart, 2013). The system of onscreen evaluation has grown more sophisticated with time. Evaluators are now able to assign scores as well as indicate remarks in digital form. There are also checks to reduce the probability of clerical errors such as ensuring marks awarded by the evaluator do not exceed the maximum marks and the tabulation of marks awarded by the system. This process has facilitated ease, speed and accuracy reducing the time required for evaluation and declaration of results. A study by Benton (2015) also found that there is no evidence of moving components to marking on-screen having any effect on the

reliability of marking; either positive or negative and limited available research has generally supported the comparability of online and traditional paper scoring. In addition, the process of onscreen evaluation has improved security as it has minimised the chance of loss of answer-books. Since raters are gathered in a single location this has also reduced subjectivity, as discussion to establish consistency in inter-rater reliability can take place.

Inception of Online Testing in India

Online assessment (i.e., testing through computers) has been in the news since the early 1990s as well-known large scale standardised assessments such as the ASVAB and GRE began to be administered through Computer Based Testing (CBT). A number of studies show that in equivalent test administration conditions there was no significant difference between the performance or results of examinees in Computer Based Tests (CBT) and Paper Pencil Tests (PPT) (e.g., Mason, Patry and Bernstein, 2001).

In India, this testing design began to take off in 2005 though on a relatively smaller scale. Demonstration of the Online Testing capability enthused various test conducting bodies in the country. This is because of the inherent potential of this form of testing to enable measurement of constructs or skills that cannot be fully or appropriately captured by paper-based tests (Bennett 2002; Parshall, Harmes, Davey, & Pashley, 2010). Another advantage is the precision or efficiency of the measurement process that this form of testing provides (Parshall, Spray, Kalohn, & Davey, 2001; van der Linden & Glas, 2000; Wainer, 1990). The third advantage is that this form of testing can be leveraged to make test administration more convenient for examinees as well as examiners- specific needs can be accommodated for presenting instructions and test items during an assessment to students with a disability. However, in the early stages in India, CBT could not be scaled up to test larger numbers of candidates in any examination. This was largely because of the infrastructure demands of the model through which these online examinations were conducted i.e., the central model (discussed later). Large scale high stakes examinations including those conducted for the award of Degrees, Diplomas and Certifications thus continued to be administered through

paper-pencil mode. A few Certification examinations conducted for smaller numbers converted to the objective format and were administered through CBT.

Conducting Online Examination through CBT Mode

Theoretically a computer can collect candidate identification data, administer and time the test, and produce a score report. A field experiment by Boevé, A. J., et al (2015) introducing Computer-Based Testing in High-Stakes Exams in Higher Education found that computer-based exam total scores are similar as paper-based exam scores, but that for the acceptance of high-stakes computer-based exams it is important that students practice and get familiar with this new mode of test administration.

In Computer Based Testing (CBT) all the candidates are given a fixed set of questions in each session of examination. Availability of the items and test for the duration of the assessment period without any testing breach or irregularity is vital. However, the administration environment for CBTs can be 'programmed' to reduce the possibility of students copying from one another through organised collusion. For example, questions and alternatives can be randomised resulting in each candidate getting questions and alternatives in a unique order in order to reduce the use of unfair means by candidates (copying from one another). Research also shows that item order does not impact student performance (ETS Report 2011).

IBPS began using CBT test design for assessments with a relatively small candidate count from 2006 onwards and in 2012 IBPS conducted its first large scale assessment for lakhs of candidates. In 2015, IBPS switched over entirely to CBT from paper-pencil testing and today it is conducting online examinations for all its client organisations (IBPS Annual Report 2017). Based on the experience and challenges which arose, the system of conducting examination in this format has been streamlined and refined to the extent that all over a crore of candidates can be tested annually and their results declared in the stipulated time.

Table showing number of candidates appearing in online tests from 2014-2019 (IBPS Annual Report 2015-2020)

Year	Recruitment		Promotion		Admission	Total
	Banking	Non-Banking	Banking	Non-Banking		
2014-15	8,127,211	1,782,286	146,310	6,848	239,683	10,302,338
2015-16	5,790,677	2,526,633	141,051	12,822	518,121	8,989,304
2016-17	7,868,427	1,805,888	127,880	13,939	848,697	10,664,831
2017-18	5,009,589	1,844,851	185,847	15,108	702,591	7,757,986
2018-19	4,738,365	1,268,014	173,710	20,827	729,721	6,930,637
2019-20	4,907,268	4,287,730	246,853	9,540	246,111	9,697,502

From Institute of Banking Personnel Selection. (2015-2020) Organisation Annual Report. Fiscal Year 2015-2020

With the success of examinations conducted by IBPS through CBT, many examination bodies switched over to the CBT model for employment as well as admission testing. This mode of testing is largely environment friendly as it completely eliminates the infrastructure requirement of question papers and answersheets in terms of printing, storage and distribution. It ensures accuracy in terms of scoring and allows for huge volumes of data to be captured and stored in digital format providing a huge reservoir of data for research. It also eliminates the time required for scanning the documents for scoring as was required with OMR technology. The results of examinations under CBT can be processed immediately. The results for even very large scale objective type examinations involving lakhs of candidates can be declared within a few days after the examination.

CBT through Central Model

Csapó, Benő, et al 2012, reviewed the contribution of new Information Communication Technologies to the advancement of educational assessment. Though intervention of this technology in assessment started on a small scale, it did not gain momentum. This is because under this model, real time internet connectivity was required for the entire duration of the examination i.e., the

systems of all candidates had to be connected to a central server through the internet at all times. Therefore, availability of continuous internet connectivity with sufficient bandwidth was required. Moreover, adequate infrastructure facilities (computer laboratories and systems) were also not available in the country to support large scale examinations through this mode. On account of these infrastructural constraints, this model could not be scaled-up. An advantage though was that the result processing could be completed in quick time.

CBT through Distributed Model

In 2011, IBPS began conducting online examinations through a distributed model. Under this model, candidates would take the examination on systems connected through a local network. In this model, internet connectivity was primarily required at the start and at the close of the examination and only intermittently during the examination. This model had greater success and within a year, IBPS could scale up from testing 200 candidates simultaneously in a session to 1,60,000 candidates simultaneously in a single session throughout the country.

Conducting Online Examinations through CAT

One practical consideration with regard to the use of CBT for large scale assessments is that unlike in paper-pencil administration where a relatively fewer sessions of examination are required owing to the availability of infrastructure, test administration through CBT often spreads over 'windows' covering days or weeks. Moreover, in the interests of test security, items which are pre-tested are rarely used for large scale assessments. Hence the test development requirement is higher for CBT. According to an ETS Report (2011) an adaptive test can match the precision of a conventional test containing 25% more items as unlike in conventional forms of CBT discussed earlier, Computer Adaptive Testing (CAT) has the capability of tailoring itself to the ability of the candidate being tested.

In the late 1960s, CAT emerged as a new field for testing after research study conducted based on Item response theory (IRT) and Adaptive testing by Frederic Lord (Wieiss, 2004) under the Personnel and Training Research Programs of

the U.S. Office of Naval Academy. CAT is the redesign of psychological and educational tests for effective and efficient administration by interactive computers. Its objective is to select, for each examinee, the set of test questions that simultaneously most effectively and efficiently measure that person on the trait (Van der Linden & Glas, 2000; Wainer et al., 2000). CAT builds on and improves upon Binet's implementation of adaptive testing by replacing the human administrator with a computer programme.

In CAT, the examinee takes the test while sitting at a computer terminal. Every candidate is administered a few questions at the start. Based on the responses given by the candidate, the system estimates the ability level of the candidate. The next set of questions is administered to the candidates matching their estimated ability level. In general, an examinee who answers a subtest item correctly will receive a more difficult item, whereas an examinee who fails that item will receive an easier item. The computer uses item response theory as a basis for selecting items. The examination is terminated at different times for different candidates reflecting their comparative ability level. In this system, candidates with higher ability level need not answer too many easy questions and candidates with lower ability level are not exposed to too many difficult questions. Thus, the system optimises the test taking for different candidates (Weiss, 2004). This system however requires that all the test items should have the parameter value obtained in the actual examination. The item bank of such tested items should hence be very large.

In 1979, the U.S. Department of Defense initiated a joint-Service project to develop and evaluate the feasibility of implementing a computer-adaptive version of the ASVAB. After decades of extensive research, the CAT-ASVAB was implemented operationally in 1996-1997. It was the first large-scale adaptive test battery to be administered in a high-stakes setting. In India, this form of assessment is in the nascent stage and CAT is not being implemented so far for large scale assessments especially high-stakes ones. However, in the future, examinations may be primarily conducted using this mode. The decision however will depend on balancing practical considerations of cost and item development requirements

on one hand and efficiency security and complexity and precision of assessment on the other.

Remote Proctored Examinations

Due to onset of Covid-19 in early 2020 there was a pressing need to conduct examinations via a mode in which the students were not required to step out of their homes. This gave impetus to testing through remote proctoring. In examinations conducted through this mode candidates can take the tests from computer systems at their homes. As the name itself implies, invigilation is done by proctors (invigilators) who are not in the same room as the candidate i.e., test sessions are monitored by invigilators at a central location and also through artificial intelligence tools. In remote proctored assessments most often the entire testing session for every candidate is recorded. Efforts are made to replicate the test administration experience available in a designated test centre i.e., test length, format, scoring are all identical.

One challenge for this mode is with regard to infrastructure. For remote proctored examinations there are precise requirements prescribed with regard to configuration of computer equipment (including the availability of microphone, in-built camera, browser and software versions etc.) as well as test environment. Candidates are made aware of these prior to the examination. These are also verified to ensure the testing experience is as smooth as possible. Availability of suitable devices and systems with all the students and also internet bandwidth remained a problem in the way of achieving wider and inclusive reach. Remote proctored assessments do require an investment of time on the part of candidates to ensure proper test administration. The high chance of use of malpractices casts doubts on the genuineness of scores and validity of the assessment.

Prior to the onset of the pandemic, such models were in limited use in developed countries and usually made available in addition to the test being held at physical test centres. Remote proctored assessments were also available to some extent in India generally managed by multi-national assessment companies. The sudden spurt in demand resulted in many organisations making the transition

and conducting educational and professional assessments through this mode. At present, these are essentially conventional CBT conducted 'at home'. IBPS has conducted remote proctored assessments for examinations with smaller candidate populations. However, given the challenges to ensuring the necessary rigours of test security and administration that should characterise high stakes assessments, the system, is not yet robust enough to apply to high stakes large scale assessments.

Automated Evaluation of Constructed Response Items: A Game Changer

One of the greatest problems in constructed-response testing is the time and expense involved in scoring. The scoring process requires substantial investment of time from highly trained scorers and often includes elaborate systems for monitoring the consistency and accuracy of the scores. In recent years, researchers have made a great deal of progress in using computer algorithms to score the responses. Automated scoring offers the possibility of greatly decreasing the time and cost of the scoring process, making it practical to use constructed response questions in testing situations where human scoring would be impractical or prohibitively expensive (Livingston 2009).

During the last year, IBPS embarked on the development of an algorithm for the scoring of constructed responses for test of English Language. This algorithm will not only help eliminate inherent biases and bring transparency to an intuitive process but also speed up the entire process of declaration of result with absolute accuracy. This intervention which aims to eventually leverage Artificial Intelligence and Machine Learning processes will be a game changer in evaluation of descriptive papers in the future. It has the potential to make constructed-response test questions practical for use in situations where scoring by human scorers is not a practical possibility. Moreover, the use of algorithms or Artificial Intelligence (AI) can provide insight into the interaction of examinees with the constructed response task as well as the response process as a whole. However suitable statistics must be applied to validate the scoring accuracy and tasks need to be specifically designed keeping AI scoring in mind. Practical

considerations of cost and nature of assessment must be taken into account to determine whether evaluation is to be done by human raters, an algorithm or Artificial Intelligence.

Large Scale Common Examinations

In the context of high stakes large scale testing, selection is often carried out on the basis of a single (common) assessment. This implies that accountability should be inherent in the assessment. Hence from time to time, various expert committees, have been constituted for the design of such assessments for which lakhs of candidates apply and appear. Selection is done through single tier or two-tier examinations. Various decisions such as the use of negative marking, scheme of scoring and evaluating descriptive papers are psychometric considerations which should be taken into account while designing large scale assessment systems. Technological and statistical advances as well as advancements in psychometrics have provided a number of administrative and assessment advantages in the design of these systems for educational and employment testing.

In 1969, fourteen banks were nationalised. Accountability to the government and people at large necessitated the design of a personnel selection system which would not only be fair but be perceived as such. To man the rapid branch expansion programme especially in rural and semi-urban areas, thousands of candidates applied and had to be screened in a speedy and objective manner. The sheer number of applications posed a formidable challenge for the individual organisations to handle effectively and speedily. A task force comprising bank executives, academicians and the faculty of NIBM was set up to evolve a selection methodology. The details of the system were designed by the faculty group of the Personnel Selection (PSS) Unit of NIBM. The methodology gained acceptance in the banking sector due to the involvement of stakeholders in its design, its demonstrable characteristics of speed, objectivity and fairness (Research Report- Dr. A.S. Deshpande).

In 1979, the Scheme of Common Recruitment through Banking Services Recruitment Boards (BSRBs) was introduced to streamline the personnel selection system in Public Sector Banks and from 1979 to 2000 personnel selection activities were carried out through BSRBs. In 1978-79, seven regional BSRBs were created, each for handling clerical recruitment of all the nationalised banks in its region and All- India officers' recruitment of some banks. For the State Bank of India (SBI) group, 13 Regional Rural Banks (RRBs) were established to handle clerical recruitment and the Central recruitment Board was entrusted with Officers' recruitment and co-ordinating clerical level recruitment activities of RRB-State Bank Group. In 1980 two more BSRBs were created. By 1985, the government reorganised the board into 15 BSRBs, with each BSRBs handling clerical recruitment of the public sector banks in its region and officers' recruitment for participating banks. BSRBs held meetings with the heads of regions/zones of all banks in its jurisdiction for sorting out problems related to finalising of schedules to absorption of selectees and feedback about their suitability.

From 1979, earlier as the PSS unit of NIBM and from 1984 as an autonomous, independent institute, IBPS provided selection services to BSRBs for recruitment in terms of conducting orientation for test administration and scoring methodologies, providing test materials, conducting examination and evaluation (IBPS, Research Report). Moreover, given the massive annual recruitment through BSRBs continuous psychometric research could be undertaken by IBPS to examine the examinations and build up an item bank to improve the quality of future examinations.

The concept of a Common Examination system was also adopted by many institutes in India for the purpose of admission and recruitment.

Another well-known common examination is the Common Admission Test (CAT) for admission to management courses (postgraduate and Fellow programmes) at the Indian Institutes of Management (IIMs). These institutes of national importance (designated by the Indian Institute of Management Act 2017) were established from 1961 onwards. As the number of IIMs grew, IIMs instituted this common test for admission. The examination which was first held in 1984 has

undergone several changes over the thirty-eight years since its inception. These include changes in test pattern as well as mode of delivery. The examination which was administered as paper-pencil test was administered through online mode (CBT) from 2009 onwards. Each year, an individual IIM is entrusted with the responsibility of conducting the examination. The CAT is considered an examination which essentially tests higher order skills to help identify candidates who possess 'non-universal skills' and can cope with rigorous course requirements in the field of management. IIMs then conduct their individual processes of selection for candidates shortlisted through the CAT examination. Lakhs of candidates apply for the examination (2.3 lakh candidates applied in 2021). Today the score obtained is also used as a basis for shortlisting for admission to postgraduate and doctoral courses in management by other non-IIM member institutions across the country.

An effort was made by the Government of Maharashtra to institute a Common Entrance Test (CET) for admission to Postgraduate Management Courses in Maharashtra. Subsequently a centralized admission process was introduced for admission to fulltime undergraduate and postgraduate courses in Engineering and Technology, Pharmacy and Hotel Management in Maharashtra. Today the centralized admission process is carried out by the Admissions Regulating Authority and State Common Entrance Test Cell established by the Government of Maharashtra in 2015.

Over the years, educational bodies such as the Central Board of Secondary Education introduced common examinations such as the All-India Higher Secondary Examination in 1964, as well as various national level examinations for admission to undergraduate professional courses such as medicine, engineering, teaching etc. The first Pre-Dental and Pre-Medical (All India Pre-Medical Pre-Dental Examination-AIPMT) now known as National Eligibility cum Entrance Test (NEET) was held in 1988 for approximately 70,000 candidates across all States/UTs except two. The Board also conducted various common entrance examinations such as the AIEEE (now known as Joint Entrance Engineering Examination-JEE Main) as well as NET (UGC National Eligibility Test), CTET (Central Teachers Eligibility Test) as well as recruitment examinations for Kendriya Vidyalaya Sangathans

(KVS) and Navodaya Vidyalaya Samiti (NVS) schools. From 2013 onwards, while the JEE Main examination was held for admission to IIITs, NITs and Central Funded technical Institutes (CFTIs), the JEE Advanced examination was held for admission to any of the Indian Institutes of Technology (IITs). The JEE Main was conducted by CBSE and JEE Advanced by IITs. These decisions were implemented based on the recommendations of an expert committee to rationalize the process and better evaluate talent.

With a view to simplify examination admission processes in line with National Policy an effort was also made by the Government of India to hold a common examination for undergraduate courses in India. In 2017, with the approval of the Government of India the National Testing Agency (NTA) was created to conduct the common entrance examinations conducted by CBSE including JEE-Main and NEET as well as conducting Central Universities Common Entrance Test (CUCET) for admission to different programmes of participating central universities. The effort was to standardise entrance examination processes for an estimated 40 lakh students. In addition, the move was to ensure a computer-based test design to ensure validity and security as well as increase access, to raise literacy levels and to increase the quality of admissions to higher education.

These efforts to hold common examinations or centralise admission processes were with the aims of reducing burden on students, having examination structures that tested skills deemed relevant to the courses by experts, reduce dependence on coaching as well as the time taken for declaration of results as well as to handle test administration and result declaration for lakhs of applicants.

Common Examination Concept for Participating Banking Organisations

After BSRBs, from 2000 to 2010 recruitment for entry level Officers and clerical posts was carried out by individual banks. In 2010, the proposal by IBPS to conduct Common Examination for recruitment in various Public Sector Banks was approved by Government of India. This was done to address multiplicity of efforts by both individual banks and candidates as candidates had to apply to

each bank separately. It sought to streamline the process of recruitment in terms of reducing fee expenditure for candidates as well as addressing the issue of the same set of candidates being selected in multiple banks. IBPS conducted the first Common Written Examination (CWE) in 2011 and provided standardised scores to candidates to enable them to apply to individual banks. In the maiden CWE, the interview process was conducted by the individual banks. This did not address the issue of selection of the same set of candidates by multiple banks.

Since 2012, IBPS has been carrying out the entire process for recruitment (application, two-tier examination, interview (as applicable) and allotment) in PSBs. The scheme was successful and was extended to RRBs with the required number of participants to each bank based on merit-cum-preference. The entire process is carried out through digital platforms right from application to testing to ensure access and objectivity and reducing the lifecycle of the recruitment process ensuring that banks had manpower within the stipulated time and candidates too were able to get jobs. The Common Recruitment Process (CRP) has completed over a decade of successful operation with challenges but without any glitches (IBPS Annual Report).

Till 2012, most of the examinations conducted by IBPS and also by other examining bodies were conducted through paper-pencil tests. The examinations had to be primarily held in about eighty to ninety cities in India which were connected by air, given the logistic arrangements to be made for dispatch and receipt of test material at centres. In case of remote areas there was a risk in terms of secure transportation of test material. After 2012, examinations were conducted through CBT on a regular basis. These examinations could be and were conducted in more than 200 cities and towns across the length and breadth of the country enabling participants to participate in the process at a place closer to their residence. More importantly since 2009, the application registration process was also conducted in digital form eliminating the requirement of candidates sending the physical application form to the Post Office resulting in delay and loss of applications. The practice of dispatching call letters in physical mode was discontinued owing to the non-receipt of call letters by a sizeable number of candidates. The process of physical dispatch was replaced with the facility to

download call letters. Through digital platforms, reach and access improved as candidates could register in the country, make fee payment online through digital mode and also download their call letter and take printout for appearing in the examination from anywhere (IBPS, Research Monograph,2016). This enabled the participants living in remote areas of the country to participate in a national level recruitment process. A study by (Konar, 2016) found that in the Common Written Examination of Probationary Officers/Management Trainees, participation of candidates from the rural and semi-urban areas increased substantially with the introduction of digital platforms. Moreover, 65 percent of candidates who were provisionally allotted to Participating Organisations belonged to places other than metropolitan cities and state capitals.

National Recruitment Agency (NRA)

The Government of India has set up National Recruitment Agency (NRA), an independent body to conduct a common eligibility test for aspirants with educational qualifications of Std. X, Std. XII and Graduation for the identified jobs in Group B and below positions in Government departments, Public Sector Undertakings and also in State Governments as well as State Government Organisations. NRA is expected to conduct its first set of examinations in the next year. The examination will be held to shortlist candidates to participate in the subsequent selection processes namely second tier of examination and interview (as applicable).

Covid-19 Fallout for Educational Institutes

With the onset of the pandemic, academic institutions faced huge challenges in terms of delivery of education as well as assessment. As educational institutions had to remain shut examinations could not be held at physical locations as done traditionally. Many Universities experimented with the conduct of CBT through learning management systems and varied digital devices including mobile phones, and iPads in addition to traditional computer systems i.e., laptops and desktops. Questions were displayed to students onscreen and the student was required to type out their answers and upload to a designated site/ link provided

by the institution. It was difficult to ensure that the rigours of examinations namely confidentiality, reliability and validity were met. However, the experience from this experiment can be leveraged to create a more robust interface for conducting such examinations in online mode so that students unable to travel to examination centres are not left out of assessment.

Evaluation Anxiety and Implications for Testing

One of the crucial aspects in a discussion on testing, is that an individual's interests and motivations interact with the testing situation and may impact test performance. During any task performance both self-relevant and task-relevant variables interact. There are a variety of self-relevant factors which have the potential for adversely impacting performance on ability assessment especially in evaluative and stressful conditions. In order to ensure the validity of assessment, the consequences of the constructs of aspiration (setting of goals relative to a person's ability and past experience), need for achievement (which includes both a motive to achieve success or a motive to avoid failure) and test anxiety (tension and apprehensiveness associated with taking a test, frequently resulting in a decrease in test performance) on test performance must be considered (Stricker, 2013; Powers, 1986; Lefcourt, 1982; Phares, 1976).

Powers (1986, 1988) found moderate negative correlations between a test anxiety measure and scores on GRE General Test. He further found that worrisome thoughts such as doing poorly on the test interfered with concentration on the test. In another study (Powers, 2001), used a newly introduced computer adaptive version of the test used in the earlier study and found that the closer match between the ability of the test-taker and item difficulty, provided by the computer adaptive version did not markedly reduce test anxiety.

Test anxiety can also be understood as arising out of the need to avoid failure and in combination with defensiveness, risk taking and creativity. Kogan and Wallach (1967) found that group discussion could enhance the risk-taking level of the group relative to the members' initial level of risk taking. Their research revealed that test anxious groups fearful of failure tend to diffuse responsibility

to reduce the possibility of personal failure and defensive groups interact insufficiently for the shift in risk taking to occur. Similarly, those with low anxiety and low defensiveness were found to take greater risks (Alker, 1969).

Some research has also shown a gender difference in the relationships of test anxiety, defensiveness and creativity (Wallach & Kogan, 1965) i.e., for boys, defensiveness was related to creativity but test anxiety was not-the more defensive were less creative); whereas for girls neither variable was related to creativity. Research also shows that cognitive performance of defensive test-takers is impaired in ambiguous contexts. Klein et al (1969) reported a U-shaped relationship between a test anxiety measure and two creativity measures and speculated that the low anxious participants make many creative responses because they do not fear ridicule, high anxious participants make many poor-quality responses because they fear a low score on the test and the middling anxious participants make few responses because their two fears cancel each other out.

Similarly, research on the level of aspiration or need for achievement and academic performance yields mixed results; with some research (Schultz and Ricciuti, 1954) showing that level of aspiration measures based on general ability test, a code learning task and regular college examinations did not correlate with college grades. On the other hand, research by Myers (1965) revealed that a questionnaire measure of achievement motivation had a substantial positive correlation with high school grades.

Putwain (2008) cautions that policy initiatives which seek to increase the frequency and importance of testing, and the testing of children at younger ages for use as performance indicators of schools and teachers do result in an increase in assessment-related anxiety in children. Hembree's 1988 meta-analysis of 562 criteria-selected studies offered statistical evidence supporting the rise of Test Anxiety as a major psychological construct impacting education. According to his findings "IQs, aptitudes, and progress of test anxious students are consistently misinterpreted and undervalued, which systematically bias the entire testing process". He further determined that "improved test performance

and Grade Point Average (GPA) consistently accompany test anxiety reduction, with an improvement of about 6 points on a 100-point scale based on 107 of 120 test anxiety treatment studies".

Test anxiety is a legitimate concern in all high-stakes testing contexts with implications for both test-takers and examination bodies or organisations undertaking selection in terms of increase in error of measurement. Moreover, when new kinds of tests and delivery systems are introduced, the implications of test anxiety (whether as a personality trait, an emotional state or a generalized anxiety disorder) must be kept in mind. Familiarity with modes of testing cannot be taken for granted and test administration bodies must be vigilant to control to the extent possible that bias (overestimation or underestimation of scores) is minimised as not all uncertainties and conditions can be controlled for. In addition to interventions to alleviate test or evaluation anxiety such as academic skill building, cognitive restructuring, relaxation strategies etc.; certain measures to address and minimize sources of anxiety arising from assessment specific features can be inbuilt by policymakers while conceptualizing the assessment methodology, design and administration process.

Preparing for the Future

Undoubtedly, technology is a facilitator and enhances the efficiency, speed and accuracy provided we have robust, reliable and valid tools / test items for testing and assessment. Leveraging new digital technologies and Artificial Intelligence and Machine Learning efficaciously and appropriately depends on human competency and ability to design systems as per the need. According to the UNESCO's 2021 Recommendation on the Ethics of Artificial Intelligence: "AI (Technology) has indeed the potential to radically reduce inequalities, promote diversity and benefit humanity as a whole, provided national and international policies as well as regulatory frameworks ensure that human-centred technologies benefit the greater interest." In light of future economic, technological and social developments (demography, changes in education and skills as per requirement, privacy and security considerations) that will influence learning and assessment, immediate practical administrative choices that lie

before the examining bodies conducting large scale examinations especially high-stake ones where accountability and validity are crucial will be as follows –

- Carry out evaluation of paper and pencil tests through computerised means at specified or centralized location(s)
- Conduct Internet based remote proctored examinations for less critical assessments in which the students be permitted to take their examinations from home.
- Conduct high stake examinations in specified centres under CBT either purely in objective format or in a combination of objective and descriptive (constructed response) format.
- Carry out online evaluation of responses, keyed in by candidates in examinations conducted through CBT, by human evaluators in specified centres.
- Carry out automated evaluation of responses typed by candidates applying AI/ML.

Conclusion

Testing and Assessment are irreplaceable and a vital component for understanding, diagnosing and predicting individual differences for different purposes and in different settings. In high stakes contexts, test scores will remain integral for decision making for the foreseeable future. The challenge will be to interpret scores appropriately for all examinees. Meeting this challenge will require mindful awareness of the factors that may cause scores to be lower than they should be for some individuals. Test Anxiety is prevalent enough and severe enough especially in high stakes assessment contexts and deserves due attention when designing assessment systems and methodologies.

With the advancement in psychometric theory and technology, various means to ensure testing and assessment is more interactive, accessible, scientific, secure, precise, accurate and robust can be explored and achieved. Technology makes it easier to reframe test development, design and administration procedures. To make selection through large scale assessments more efficient, it is advisable to devise a system in which basic skills assessment could be carried out through modes which afford control to a baseline acceptable degree and higher order skills assessment which comprises the major weightage for selection should be carried out without sacrificing any of the rigours of the assessment process. Research and experimentation based on pilot studies are a must to ensure that systems are used to enhance the capability to store and process data in order to make better informed decisions in selection and make them more engaging and interactive. Examination bodies across the board will also need to invest in developing and integrating processes alongside implementation of Artificial Intelligence and Machine Learning to ensure rigour in conducting examinations known as Internet Based Examination or Remote Proctored Examinations. Finally, feedback from stakeholders as well as shared insights among psychometricians and various examination conducting institutions are a must for devising and framing a robust successful process and policy for conducting examinations.

References

Alker, H. A. (1969). Rationality and Achievement: A comparison of the Atkinson-McClelland and Kogan-Wallach formulations. Journal of Personality, 37, 207–224. https://doi.org/10.1111/j.1467-6494.1969.tb01741.x

Angoff, W. H. (1953). Test reliability and effective test length. Psychometrika, 18 (1), 1-14.

Attali, Y, and Sinharay. S. (2015). Automated Trait Scores for TOEFL® Writing Tasks. ETS Research Report Series, no. 1, 1-14.

Attali, Y. (2015) Reliability-based feature weighting for automated essay scoring. Applied Psychological Measurement 39, no. 4 (2015): 303-313.

Bennett, R. E. (2002). Using electronic assessment to measure student performance: Online testing. State Education Standard, 3 (3), 23-29.

Benton, T. (2015). Examining the impact of moving to on-screen marking on concurrent validity. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Binet A. (1905). The Development of Intelligence in Children (The Binet-Simon test) trans. Kite Elizabeth. Baltimore, MD: Williams and Wilkins.

Boevé, A. J., Meijer, R.R., Albers, C.J. Beetsma, Y. and Bosker, R.J. (2015). "Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment." PloS one 10, no. 12 (2015): e0143616.

Bonaccio, S., & Reeve, C. L. (2010). The nature and relative importance of students' perceptions of the sources of test anxiety. Learning and Individual Differences, 20 (6), 617-625.

Brown, Gavin. (2019). "Is assessment for learning really assessment?" In Frontiers in Education, vol. 4, p. 64. Frontiers, 4.10.3389/feduc.2019.00064.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A. M., & Kaemmer, B. (1989). Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring. Minneapolis: University of Minnesota Press

Byrne, D. G., Davenport, S. C., & Mazanov, J. (2007). Profiles of adolescent stress: The development of the adolescent stress questionnaire (ASQ). Journal of adolescence, 30 (3), 393-416.

Camara, Wayne J., and Gary Echternacht. (2000). The SAT [R] I and High School Grades: Utility in Predicting Success in College. Research Notes.

Chandio, M. T., Pandhiani, S. M., & Iqbal, R. (2016b). Bloom's taxonomy: Improving Assessment and Teaching-Learning Process. Journal of Education and Educational Development, 3(2), 203. https://doi.org/10.22555/joeed.v3i2.1034

Chandio, Muhammad Tufail, Saima Murtaza Pandhiani, and Rabia Iqbal. (2016). "Bloom's taxonomy: Improving assessment and teaching-learning process." Journal of Education and Educational Development 3, (2) .

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. Educational measurement: Issues and practice, 20 (4), 19-27.

Cohn, Elchanan, Sharon Cohn, Donald C. Balch, and James Bradley Jr. (2004). "Determinants of undergraduate GPAs: SAT scores, high-school GPA and high-school rank." Economics of education review 23 (6), 577-586.

Coyle, Thomas R., and David R. Pillow. (2008). "SAT and ACT predict college GPA after removing g." Intelligence 36 (6), 719-729.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cronbach, L. J. (1990). Essentials of psychological testing (4th edn) (New York, Harper & Row).

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2011). Technological issues for Computer-Based assessment. In Springer eBooks (pp. 143–230). https://doi.org/10.1007/978-94-007-2324-5_4

Csapó, Benő, John Ainley, Randy E. Bennett, Thibaud Latour, and Nancy Law. (2012). Technological issues for computer-based assessment. Assessment and teaching of 21st century skills 143-230.

Davey, T., & Lee, Y. H. (2011). Potential impact of context effects on the scoring and equating of the multistage GRE® revised General Test. ETS Research Report Series, 2011(2), i-44.

Deary, I. J. (2001). Intelligence: a very short introduction (New York, Oxford University Press).

Deng, Hui, Feng Wang, and Bo Liang. (2008) A low-cost OMR solution for educational applications." In 2008 IEEE international symposium on parallel and distributed processing with applications, pp. 967-970. IEEE, 2008

Derakshan, N., & Eysenck, M. W. (2009). Anxiety, processing efficiency, and cognitive performance: new developments from attentional control theory. European Psychologist, 14 (2), 168-176.

Dolke, A.M. (1978). GATB- Aptitude structure and Norms for general working population and specific jobs in Textile Industry [Unpublished doctoral dissertation]. Gujarat University.

Deshpande, A.S. (1986). Large scale selection examination: Conceptual considerations and operational strategies with human systems and modern technology [Report submitted to Indian Association of Educational Technology, New Delhi].

Deshpande, A.S. (1988). Predictive Validity of Selection Tests and Interviews in Personnel Selection: A review, [Unpublished Manuscript] IBPS Report Repository.

Dubois, T. E., & Cohen, W. (1970). Relationship between measures of psychological differentiation and intellectual ability. Perceptual and Motor Skills, 31 (2), 411-416.

Efendi, Raimon, Lido Sabda Lesmana, Firmansyah Putra, Efri Yandani, and Ratih Agustin Wulandari. (2021). Design and Implementation of Computer Based Test (CBT) in vocational education. In Journal of Physics: Conference Series, 1764 (1), p. 012068. IOP Publishing.

Freeman, Frank S. (1962). "Theory and practice of psychological testing. New York: Holt." Rinehart, Winston

Goslin, David A. (1963). The search for Ability: Standardized Testing in Social Perspective. Vol. 1. Russell Sage Foundation.

Gregory, Robert J. (2014). Psychological testing: History, principles, and applications. Pearson Education Limited.

Gulliksen, Harold. (1950). Intrinsic validity. American Psychologist 5 (10), 511.

Heiman, T., & Precel, K. (2003). Students with learning disabilities in higher education: Academic strategies profile. Journal of learning disabilities, 36 (3), 248-258.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. Review of educational research, 58 (1), 47-77.

Howard, E. (2020). A review of the literature concerning anxiety for educational assessments. Published by Ofqual, UK 1-63.

Hubin, David Royce. (1988). The Scholastic Aptitude Test: its development and introduction, 1900-1948. [Unpublished doctoral dissertation]. University of Oregon.

Hurley, M., & Padró, F. F. (2006). Test Anxiety and high stakes testing. International Journal of Learning, 13 (1), 163-170.

Hunter, J. E. (1983). The Dimensionality of the General Aptitude Test Battery (GATB) and the Dominance

of General Factors over Specific Factors in the Prediction of Job Performance for the US Employment Service. Michigan State Dept. of Labor, Detroit. Michigan Employment Security Commission. https://eric.ed.gov/?id=ED236166

Institute of Banking Personnel Selection (2011). Organisation Annual Report. Fiscal Year 2011.

Institute of Banking Personnel Selection (2017). Organisation Annual Report. Fiscal Year 2017.

Klein, S. A., Frederiksen, N., & Evans, F. R. (1969). Anxiety and learning to formulate hypotheses. Journal of Educational Psychology, 60(6, Pt.1), 465–475. https://doi.org/10.1037/h0028351

Klein, S. P., Frederiksen, N., & Evans, F. R. (1969). Anxiety and learning to formulate hypotheses. Journal of Educational Psychology, 60, Number 465–475. https://doi.org/10.1037/h0028351

Klinger, D. A., Freeman, J. M., Bilz, L., Liiv, K., Ramelow, D., Sebok, S. S., Samdal, O., Dür, W., & Rasmussen, M. (2015). Cross-national trends in perceived school pressure by gender and age from 1994 to 2010. European Journal of Public Health, 25(suppl 2), 51–56. https://doi.org/10.1093/eurpub/ckv027

Klinger, D. A., Freeman, J. G., Bilz, L., Liiv, K., Ramelow, D., Sebok, S. S., … & Rasmussen, M. (2015). Cross-national trends in perceived school pressure by gender and age from 1994 to 2010. The European Journal of Public Health, 25 (suppl_2), 51–56.

Kogan, N., & Wallach, M. A. (1967a). Effects of Physical Separation of Group Members upon Group Risk-taking. Human Relations, 20(1), 41–48. https://doi.org/10.1177/001872676702000104

Kogan, N., & Wallach, M. A. (1967b). Group risk taking as a function of members' anxiety and defensiveness levels. Journal of Personality, 35(1), 50–63. https://doi.org/10.1111/j.1467-6494.1967.tb01415.x

Kogan, N., & Wallach, M. A. (1967). Effects of physical separation of group members upon group risk taking. Human Relations, 20, Number 41–49. https://doi.org/10.1177/001872676702000104

Kogan, N., & Wallach, M. A. (1967). Group risk taking as a function of members' anxiety and defensiveness levels. Journal of Personality, 35, Number 50–63. https://doi.org/10.1111/j.1467-6494.1967.tb01415.x

Kogan, N., & Wallach, M. A. (1967c). Risky-shift phenomenon in small decision-making groups: A test of the information-exchange hypothesis. Journal of Experimental Social Psychology, 3(1), 75–84. https://doi.org/10.1016/0022-1031(67)90038-8

Kogan, N., & Wallach, M. A. (1967). Risky-shift phenomenon in small decision-making groups: A test of the information-exchange hypothesis. Journal of Experimental Social Psychology, 3 (1), 75-84.

Konar, AK. (2016). Recruitment of Entry level Officers and Clerks in Public Sector Banks (PSBs) –Profile of candidates with regard to their place of Residence. IBPS Research Monograph 1 (1), 1-9.

Lefcourt, H. M. (1982). Locus of control: Current trends in theory and research (2nd ed.). Hillsdale: Erlbaum. In Stricker, L.J. (2013), ETS RESEARCH ON COGNITIVE, PERSONALITY, AND SOCIAL PSYCHOLOGY: I. ETS Research Report Series, 2013: i-29. https://doi.org/10.1002/j.2333-8504.2013.tb02308.x

Linden, Wim J., Wim J. van der Linden, and Cees AW Glas. (2000) (Eds). Computerized adaptive testing: Theory and practice. Springer Science & Business Media.

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding Their Impact. Journal of Applied Testing Technology, 13(3).

Lissitz, Robert W., Xiaodong Hou, and Sharon Cadman Slater. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. Journal of Applied Testing Technology 13 (3), Page numbers Retrieved from https://jattjournal.net/index.php/atp/article/view/48366

Livingston, S. A. (2009). Constructed-Response Test Questions: Why We Use Them; How We Score Them. R&D Connections. Number 11. Educational Testing Service.

Maier, M. H. (1993). Military aptitude testing: The past fifty years (Tech. Rep. No. 93-007). Seaside, CA: Defense Manpower Data Center.

Martinez, M. E., & Bennett, R. E. (1992). A review of automatically scorable constructed-response item types for large-scale assessment. ETS Research Report Series, 1992 (2), i-34.

Mason, B. Jean, Marc Patry, and Daniel J. Bernstein. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. Journal of Educational computing research 24 (1), 29-39.

McIntire, Sandra A.; Miller, Leslie A. (2006). Foundations of Psychological Testing: A Practical Approach, Second Edition. SAGE Publications (CA)

McClellan, C. A. (2010). Constructed-response scoring—Doing it right. R&D Connections, 13, 1-7.

McCrae, R. R., & Costa Jr, P. T. (1995). Trait explanations in personality psychology. European Journal of Personality, 9 (4), 231-252.

Moreno, Kathleen E., C. Douglas Wetzel, James R. McBride, and David J. Weiss. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. Applied Psychological Measurement, 8 (2), 155-163.

Myers, A. E. (1965). Risk Taking and Academic Success and their Relation to an Objective Measure of Achievement Motivation. Educational and Psychological Measurement, 25(2), 355–363. https://doi.org/10.1177/001316446502500206

Myers, A. E. (1965). Risk taking and academic success and their relation to an objective measure of achievement motivation. Educational and Psychological Measurement, 25, Number 355–363. https://doi.org/10.1177/001316446502500206

Newstead, S., & Dennis, I. (1994). The reliability of exam marking in psychology: examiners examined. Psychologist, 7(5), 216-219. Newstead, S. & Dennis, I. (1994) The reliability of exam marking in psychology: examiners examined, Psychologist, 7 Number, 216–219.

Oh, I., Schmidt, F., Shaffer, J. A., & Le, H. (2008). The Graduate Management Admission Test (GMAT) is Even More Valid Than We Thought: A New Development in Meta-Analysis and Its Implications for the Validity of the GMAT. Academy of Management Learning and Education, 7(4), 563–570. https://doi.org/10.5465/amle.2008.35882196

Oh, In-Sue, Frank L. Schmidt, Jonathan A. Shaffer, and Huy Le. (2008). The Graduate Management Admission Test (GMAT) is even more valid than we thought: A new development in meta-analysis and its implications for the validity of the GMAT. Academy of Management Learning & Education 7 (4), 563-570.

Parshall, C. G., Spray, J. A., Kalohn, J., & Davey, T. (2002). Practical considerations in computer-based testing. Springer Science & Business Media.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2009). Innovative items for computerized testing. In Elements of adaptive testing (pp. 215–230). New York, NY: Springer New York. In Davey, T. (2014). 3.1 Practical Considerations in Computer-Based Testing1. A Compendium of Studies.

Pekrun, R., Goetz, T., Perry, R. P., Kramer, K., Hochstadt, M., & Molfenter, S. (2004). Beyond test anxiety: Development and validation of the Test Emotions Questionnaire (TEQ). Anxiety, Stress & Coping, 17 (3), 287–316.

Phares, E. J. (1976). Locus of control in personality. Morristown: General Learning Press. In Stricker, L.J. (2013), ETS RESEARCH ON COGNITIVE, PERSONALITY, AND SOCIAL PSYCHOLOGY: I. ETS Research Report Series, 2013: i-29. https://doi.org/10.1002/j.2333-8504.2013.tb02308.x

Powers, D. E. (1986). Test anxiety and the GRE General Test (GRE Board Professional Report No. 83-17P). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1986.tb00200.

Powers, D. E. (1988). Incidence, correlates, and possible causes of test anxiety in graduate admissions testing. Advances in Personality Assessment, 7, 7, Number 49–75.

Powers, D. E. (1986). Test anxiety and the GRE General Test (GRE Board Professional Report No. 83-17P). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1986.tb00200.

Powers, D. E. (2001). Test anxiety and test performance: Comparing paper-based and computer-adaptive versions of the Graduate Record Examinations (GRE) General Test. Journal of Educational Computing Research, 24, 249–273. https://doi.org/10.2190/680W-66CR-QRP7-CL1F

Puente, A. E., & Garcia, M. (2000). Psychological Assessment of Ethnic Minorities. In Elsevier eBooks (pp. 527–551). https://doi.org/10.1016/b978-008043645-6/50099-9

Puente, Antonio E., and M. Perez-Garcia. "Psychological assessment of ethnic minorities." Handbook of psychological assessment (2000): 527-551.

Putwain, D. W. (2008). Deconstructing test anxiety. Emotional and Behavioural Difficulties, 13 (2), 141-155.

Putwain, D. W. & Daly, A. L. (2014). Test anxiety prevalence and gender differences in a sample of English secondary school students. Educational Studies, 40 (5), 554-570

Sands, William A., Brian K. Waters, and James R. McBride. (1997). Computerized adaptive testing: From inquiry to operation. American Psychological Association.

Schaeffer, Gary A., Manfred Steffen, Marna L. Golub-Smith, Craig N. Mills, and Robin Durso. (1995). The introduction and comparability of the computer adaptive GRE general test. ETS Research Report Series, no. 1, i-48.

Scheuneman, J. D., & Oakland, T. (1998). High-stakes testing in education. In J. H. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), Test Interpretation and Diversity: Achieving equity in assessment (pp. 77–103). American Psychological Association. https://doi.org/10.1037/10279-004

Schrader, William B. (1984). The Graduate Management Admission Test: Technical Report on Test Development and Score Interpretation for GMAT Users Educational Testing Service, Princeton, N.J.

Schultz, D. H., & Ricciuti, H. N. (1954). Level of aspiration measures and college achievement. Journal of General Psychology, 51(2), 267–275. https://doi.org/10.1080/00221309.1954.9920226

Schultz, D. G., & Ricciuti, H. N. (1954). Level of aspiration measures and college achievement. Journal of General Psychology, 51 Number 267–275. https://doi.org/10.1080/00221309.1954.9920226

Scroggins, W. A., Thomas, S. L., & Morris, J. A. (2008). Psychological testing in personnel selection, part I: a century of psychological testing. Public Personnel Management, 37 (1), 99-109.

Segall, D. O., & Moreno, K. E. (1999). Development of the Computerized Adaptive Testing Version of the Armed Services Vocational Aptitude Battery. In Innovations in Computerized Assessment (1st ed., pp. 35–65). Taylor & Francis Group. https://doi.org/10.4324/9781410602527-9

Segall, Daniel O., and Kathleen E. Moreno. (1999). Development of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery. Innovations in computerized assessment, Volume Number 35-65.

Shi, F. (2012). Exploring Students' Anxiety in Computer-based Oral English Test. Journal of Language Teaching and Research, 3(3), 446–451. https://doi.org/10.4304/jltr.3.3.446-451

Shi, F. (2012). Exploring Students' Anxiety in Computer-based Oral English Test. Journal of Language Teaching & Research, 3(3). Page number

Swart, Arthur James. (2013). Onscreen Marking: An effective assessment tool for engineering education in the information age. ICEE/ICIT-2013 CONFERNCE (2013): 489.

UNESCO. (2022). State of the education report for India, 2022: Artificial intelligence in education; here, there and everywhere. ISBN 978-81-89218-83-6

Urbina, Susana. (2014). Essentials of psychological testing. John Wiley & Sons.

van der Linden, Wim J., and Cees AW Glas. (2000). Capitalization on item calibration error in adaptive testing. Applied Measurement in Education 13 (1), 35-53.

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. Journal of affective disorders, 227, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Wainer, H., Dorans, N. J., Flaugher, R. L., Green, B. F., & Mislevy, R. J. (1990). Computerized Adaptive Testing: a primer. http://ci.nii.ac.jp/ncid/BA49621425

Wainer, Howard, Neil J. Dorans, Ronald Flaugher, Bert F. Green, and Robert J. Mislevy. (2000). Computerized adaptive testing: A primer. Routledge.

Wallach, M. A., & Kogan, N. (1965). Modes of thinking in young children—A study of the creativity-intelligence distinction. New York: Holt, Rinehart and Winston in Stricker, L.J. (2013), ETS RESEARCH ON COGNITIVE, PERSONALITY, AND SOCIAL PSYCHOLOGY: I. ETS Research Report Series, 2013: i-29. https://doi.org/10.1002/j.2333-8504.2013.tb02308.x

Weiss, David J. (2004). Computerized adaptive testing for effective and efficient measurement in counselling and education. Measurement and Evaluation in Counselling and Development 37 (2), 70-84.

White, Margaret B., and Alfred E. Hall. (1980). An overview of intelligence testing." Educational Horizons 58 (4), 210-216.

Wiggins, G., & McTighe, J. (1998). Understanding by Design. ASCD. http://www.c3schools.org/Presentations/OakHill/vignettes.pdf

Wiggins, Grant P., Grant Wiggins, and Jay McTighe. (2005) Understanding by design. ASCD.