World Conference on Research in EDUCATION
29-31 August, 2019 Berlin, Germany

Evidence Based Assessment - Policy, Principle and Practice Perspective

Vineet Joshi[1], Sadhana Parashar[2], Priyanka Sharma[3]

National Testing Agency, India

Abstract

Application of general principles of drawing conclusion based on evidence forms the foundation of Evidence-based Assessment (EBA). EBA is not a novel method or technique or model rather it is a systematic approach of constructing educational assessment based on evidence-driven reasoning from learning sciences, cognitive sciences, measurement models and other sources of evidence from assessment related disciplines. Educational assessments of all kinds try to draw reason from test takers and draw conclusions about their knowledge and abilities. Rooting these conclusions on evidence and using them to support conclusions is what forms the core of EBA. Over the years, several assessment methods, including latent trait theories like IRT have evolved to improve the validity of conclusions drawn from the test scores, in a more scientific manner. This paper discusses the significance and usage of variety of evidence from various sources, at different stages of educational testing to improve the quality of the test, especially validity of the conclusion drawn on the basis of the test scores about test takers. It is not limited to, but briefly touches the principle and practices of Evidence-Centered Design (ECD), and tries to establish linkages at various points of the testing cycle. It extends to using evidence from measurement models from Classical Test Theory and Item Response Theory. The intent is to produce usable evidence and present evidence in simpler form to users, who may be great subject experts and instruction designers or policy makers but may not be great measurement experts.

Keywords: Testing, Evidence, Psychometrics, ECD, Measurement Models

[1] Director General, National Testing Agency
[2] Senior Director, National Testing Agency
[3] Senior Research Advisor, National Testing Agency

Background

Evidence-based Assessment (EBA) is a systematic approach of constructing educational assessment based on evidence driven reasoning from learning sciences, cognitive sciences, measurement models and other importance of evidence in assessment interlinked disciplines. Educational assessments of all kinds try to draw reason from test takers and draw conclusions about their knowledge and abilities. Rooting these conclusions on evidence and using them to support conclusions is what forms the core of EBA. Over the years, several assessment methods, including latent trait theories like IRT have evolved to improve the validity of conclusions drawn from the test scores, in a more scientific manner.

One may argue – "why do you need evidence?". On several occasions, subject matter experts and other professionals get baffled by focus on evidence and debate that they have a vast experience in the field and can claim to judge abilities based on their experience. There is no denial of significance of experience and expertise. However, one needs to analyse that entire assessment practice revolves around the Dogma that some important decision is being made about someone's abilities invisible) based on their performance on a gamut of items or tasks. The central problem of test theory is the relation between the ability of the individual and his [or her] observed score on the test (Gulliksen, 1961). Much of the progress in test theory has been made by treating the study of the relationship between responses to a set of test items and a hypothesized trait (or traits) of an individual as a problem of statistical inference (Lewis, 1986).

If we look from another point, an assessment is a tool intended to illicit some intended response from students as an indicator of their behavior, in order to gather some data that can be used to draw reasonable inferences about what students know and sometimes what they don't know. This process of data or evidence collection is targeted to support the types of inferences one wants to draw, and this reasoning from evidence forms underlying principle of evidence centred assessment designs and practices. This chain of reasoning about student learning characterizes all assessments, from classroom quizzes and standardized achievement tests, to computerized tutoring programs, to the conversation a student has with her teacher as they work through an experiment (Mislevy, 1994, 1996).

One more point that we would like to make here that evidence is not only about data, because data does not provide its own meaning; it becomes evidence only if it is correlated and interpreted in a given context. Same is true for educational testing. One needs to be very careful about – "evidence of what", "which type of evidence", "how to gather evidence". Testing theories and science of measurement offer various methods for utilizing available evidence to make conclusions about the competencies of learners. Probabilistic models are one such example. It is the chain of reasoning that determines the kind of evidence required, the method and time to collect the evidence, and justification about the need of specific evidence to make

decision about students' ability. Classical Test Theories as well as Modern Test Theories leverage the underlying principles of reasoning from evidence to support inference.

The process of reasoning from evidence is deeply rooted in the Evidence-based Assessment (EBA) discussed in the paper. It is being explicitly stated here in the beginning that EBA is not a new model of theory of testing, neither it is a set of rules or procedures. Rather, it is an approach driven by reason from the evidence at various stages of testing. It may exhibit some degree of parallelism with Evidence Centred Design (ECD), as well as Principle Design Approach. However, without overestimating or undermining any of these, EBA has been adopted as a simple way of test development, implementation, scoring and interpretation of score, based on evidence from various sources. Evidence and reasons are understood (still in progress) and usable by various specialists involved in the end-to-end process of test development and delivery, including subject experts test developers, test administrators, platform/IT managers, UI developers, statisticians and psychometricians, as well as policy related decision makers.

Next sections of the paper deal with the guiding principles of ECD, approach of EBA, practical examples and experience, but before all that, it is important to understand the landscape of educational testing in India, especially high stakes examinations in the domain of Higher Education, for which this approach has been adopted to some extent.

Landscape of Educational Testing in Higher Education in India

High stakes Examinations in Indian scenario are characterised by the following features

I. Voluminous numbers and diverse cohort

On the basis of purpose, they serve, educational assessments and tests may be clubbed under three main categories:

- Entrance Examinations–Tests of these examinations are usually designed from two different perspectives - one which tests subject knowledge, achievement and learning level for the course the student has completed, another which intends to measure the knowledge and skills required for the target course. Most of the organisations of repute offer admission to the candidates solely on the basis of performance in such examinations (subject to fulfilment of minimum eligibility criteria in qualifying examination for these entrance examinations).
- Eligibility Tests–Eligibility tests differ from entrance examinations in the sense that they are one of the criteria for the defined purpose. From this viewpoint, eligibility tests fall under the category of medium stake examinations. However, due to intended/unintended usage of result, for example award of fellowship to research scholars, screening of candidates for recruitment purpose, certificate of merit associated

with the tests, stakes associated with eligibility tests are certainly high and without any exaggeration, may be considered as high stakes examination due to their profound impact on student's academic profile, career and profession.

- Recruitment Tests – These tests are conducted by an organisation for finding suitable candidates for a job in their organisation.

Authors' experience and discussion in this paper is based on first two kinds of tests. How high are the stakes associated with high stakes examinations in Indian Scenario may be inferred from Table 1 that presents an iceberg tip view of the landscape of few national level examinations.

Table 1: A Glimpse of National Examinations in the Domain of Higher Education in India

Name of the Examination	Purpose	Profile of Candidates	Number of Candidates (2018/2019 administration)
Joint Entrance Examination (JEE-Main)	Entrance to UG courses in **engineering** colleges (2 administrations in a year)	Grade 12 (Appearing or Passed)	929,198 935,755
National Entrance-cum-Eligibility Test (NEET-UG)	Entrance to MBBS/BDS in **Medical** Colleges (single administration, in 11 languages)	Grade 12 (Appearing or Passed)	1,519,375
AIIMS-MBBS Entrance Examination	Entrance to MBBS/BDS Courses in All India Institutes of Medical Sciences (single administration)	Grade 12 (Appearing or Passed)	452,931
Common Admission Test (CAT)	Entrance to MBA in IIMs and other premier MBA colleges	Graduates	231,000
Graduate Aptitude Test in Engineering (GATE)	Entrance to Master's programs and Doctoral programs in in Engineering/Technology/Architecture/Science	Engineering Graduates and Science Post-graduates	781,854

National Entrance-cum-Eligibility Test (NEET-PG)	Entrance to MD/MS/PG Diploma in disciplines of Medical Sciences	Medical/Dental/Related Graduates	714,562
University Grant Commission-National Eligibility Test (UGC-NET, Humanities)	Eligibility for Asst Professor and Junior Research Fellowship (2 administrations in a year)	Postgraduates	942,419
			935,755
University Grant Commission-National Eligibility Test (UGC-NET,	Eligibility for Asst Prof and Junior Research Fellowship	Postgraduates	250,000
GMAT+GRE	For overseas higher studies	Graduates	~100,000
IELTS	For overseas higher studies	Graduates	~700,000
PTE Academic	For overseas higher studies	Graduates	~100,000
TOEFL	For overseas higher studies	Graduates	~70,000

Qualifying rate for most of the above examinations is in the range if 1-6%, and these are just a few. 4-5 million students appear for various entrance examinations per year. It also shows journey of an Indian student who aspires to pursue a course of his/her choice.

II. Transition from Paper-Pen to Computer Based Tests

Shifting to Computer Based Tests (CBT) is rather a recent move in India for a hassle free, speedier and efficient testing in high stakes examination. Till now a few examinations were being conducted in hybrid mode, when candidates had an option to choose between paper pen and computer-based test, and CBT constituted less than 20% of the total candidates. Being a new practice, on one hand it is seen as an avenue toward greater accessibility for students, while on other hand it is seen as a threat due to digital divide, especially on equity issues.

In order to ensure that no student is at an advantage or at a disadvantage due to the current examination interface, government has set up more than 4000 free-of-cost test practice centres (TPCs) for interested candidates to practice tests in simulated condition.

III. Inherent Dilemmas

Limited opportunities and huge scale, coupled with societal aspirations where every student wishes to qualify in these examinations, testing systems witness some dilemmas:

Difficulty vs test targeting- Most of the tests are extremely difficult. It raises a very pertinent question – how well targeted are the tests. An extremely difficult test may not well target the intended population.

Construct- past knowledge and skills vs required knowledge skills – Few of the tests focus on what candidates have studied in the past like JEE(Main), NEET (UG) are based on the content taught at Grade 11 and 12, whereas CAT-GATE focus on aptitude that is expected to be found in the students pursuing relevant courses. What is more appropriate or desirable in current scenario- is another major consideration.

IV. Bring Testing to the Centre Stage in Education Policies

At various occasions, need was felt to device a national machinery in order to

- avoid multiple examinations
- provide the student with opportunities to improve their performance through multiple window testing
- provide students transparent, fair, technically valid and reliable tests

And, it was recommended to set up a specialized National Testing Service as a quality control mechanism, which would organize national level tests, set norms for high stakes examinations for comparability of performance and would also conduct independent tests (MHRD, 1986, 1990; NKC, 2009). And that is how NTA has been set up in 2018 with a mandate to improve equity and quality in education by administering research based valid, reliable, efficient, transparent, fair and international level assessments.

In order to achieve its mandates and objectives, NTA believes in learner centric, and evidence driven decisions for various examination practices.

Evidence Centred Design- The Guiding Principle

Evidence Centred Design (ECD) treats assessment as a process of reasoning from the necessarily limited evidence of what students do in a testing situation to claims about what they know and can do in the real world. Mislevy, Steinberg, and Almond (1999) described ECD as a "principled framework for designing, producing, and delivering educational assessments". Later in year 2003, a comprehensive view of the rationales underlying ECD was provided by Mislevy, Almond, and Lukas (2003), According to them, ECD is based on three premises:

1. an assessment must build around the important knowledge in the domain of interest and an understanding of how that knowledge is acquired and put to use;
2. the chain of reasoning from what participants say and do in assessments to inferences about what they know, can do, or should do next, must be based on the principles of evidentiary reasoning;
3. purpose must be the driving force behind design decisions, which reflect constraints, resources, and conditions of use.

(Mislevy, Almond, and Lukas, 2003)

Review of literature around ECD indicates that it is not a set of rigid procedures. It is a set of practices that helps test developers to draw valid inferences about the test takers based on their scores on a specific test. It also helps them designing assessments which could enable them to draw valid conclusions. In other words, it ensures that the way evidence is gathered and interpreted is consistent with the purpose of the assessment. This section is primarily based on the ECD research by Mislevy (1993, 1994), Ziecky (2014) based on

In ECD the entire process of designing, developing, and using tests into five groups of activities called "layers" (Zieky, 2014):

1) Domain Analysis,

2) Domain Modeling,

3) Conceptual Assessment Framework,

4) Assessment Implementation, and

5) Assessment Delivery.

Domain Analysis - Every test measures a sample of KSAs from a specific domain. It requires identification of the relevant domain and an investigation of its characteristics and components and then analyses following:

- What KSAs are most important?
- How are they represented?
- How are the KSAs related to one another?
- How are the KSAs generally acquired and how are they used in the real world?
- What kind of work dependent on the KSAs is valued?
- How is good work distinguished from mediocre or poor work in that domain?

Domain Modeling- The domain modeling layer moves from an investigation of the relevant real-world domain to a use of selected aspects of the domain for the purpose of building an assessment argument.

Conceptual assessment framework – Conceptual Assessment Framework consists of various tools used by test developers and is much relevant for test design and construction. It includes

- student model representing gamut of relevant KSAs
- evidence model illustrating observable behaviours related to above KSAs

- task model representing description of test items/tasks, and
- assembly model illustrating norms and specifications of the test

Assessment Implementation – This layer is similar to the traditional practice of item writing and test assembly and uses task shells generated from task model for test development.

Assessment Delivery – The last layer the assessment delivery layer comprises test administration and scoring, which may be accomplished in four stages- selection of the task to be presented to the test taker, presentation of task to elicit and record test taker's response, processing of item responses, and summary scoring based on appropriate quantitative techniques.

The Practical Approach of EBA

Let us have a holistic view of traditional test development practice that has been serving the intended purpose since many years. Undoubtedly, good test developers have always tried to define the purpose of a test in a concrete and comprehensive manner first, followed by defining the construct comprising of relevant KSAs to be measured in order to meet the purpose of a test, and then to choose the best possible ways to measure those KSAs within a given context.

EBA has all the practices of above steps:

A. defining the measure or construct, generating test specifications (or blueprint) in terms of table of specifications at content level, difficulty level, cognitive level, item format level,
B. deciphering test specifications at item level table of specification, measuring for each head under A
C. creating items by a pool of experts against specifications mentioned in B, at the appropriate levels of difficulty, cognitive process, etc
D. providing keys and rationale for distractors (as we are currently using only MCQs)
E. assembling test forms to meet specifications (for fixed test forms), moderating if needed to establish equivalence and comparability
F. contemplating response templates as per the logic for CBT, so that meaningful score could be generated

Looking at the steps, one can easily interpret that EBA is neither a substitute of any existing method nor a novel method. It is an approach that guides decisions taken at various steps of testing cycle by various personnel involved in the process. It is the common USP that influences the test design explicitly and links the decisions made during test design and development, administration, scoring, and reporting into a chain of evidence-based reasoning that supports the validity of the conclusions made about test takers on the basis of their scores. It is a supplementation to enhance efficiency, and effectiveness of traditional test development, which has been discussed in further parts of this section.

Authors wish to communicate here that the studies or subprojects discussed here are exploratory in nature and whatever indicatives results are there are yet to be established over the years. These studies have been undertaken to validate /refute certain assumptions, based on rigorous evidence, with the single purpose to improve the test quality and functioning.

Reason with Evidence: From test design to decision (What do we measure)

Learner centricity drives the construct related decision. As stated earlier, decision about students' level of ability in that construct based on performance on the test items is made by a group of experts. The test developers themselves are not expected to become expert in every domain in which they work. They are, however, expected to know how to elicit the necessary information to complete the domain analysis.

Assumption I: Experts have a fair understanding of image of candidature

Assumption II: Experts have a fair understanding of how equivalent test forms look like

As test developers and administrators produce certain claims based on test score i.e performance of test takers on a test, the first assumption is experts engaged in test development have an idea of image of candidature in terms of - what the intended target audience knows and can do in terms of KSAs; what degree of KSAs would be reflected by high performers and low performers on the test.

Evidence from various sources using both qualitative and quantitative techniques including CTT and IRT measurement models, based on tests and test takers responses from previous administrations (as baseline):

A. **Evidence from Content analysis** – In educational set ups, content analysis is considered as an appropriate method, sitting at the juncture of the qualitative and quantitative traditions. Based on content analysis of three major tests mentioned in Table 1, we found that tests do not adhere to defined table of specifications in terms of:
    a. Difficulty level
    b. Cognitive process

As purpose of the discussion is to highlight the approach and not the details of findings, the degree of deviation is not being discussed here.

B. **Evidence from Expert's perception –** In a definite design, experts were sent identified number of test forms so that **each** form receives equal number of ratings. Experts were asked to judge-
    (a) Are the test forms comparable (1- not at all; 5- absolutely comparable)? It can be seen that more than 60% experts rated 4 or 5 on comparability parameter, however approximately 40% rated that forms have either average or poor comparability Although, there are standard equating procedures in place, to establish equivalence of scores, but

considering high stake nature of examinations, comparability of test forms is an essential task.

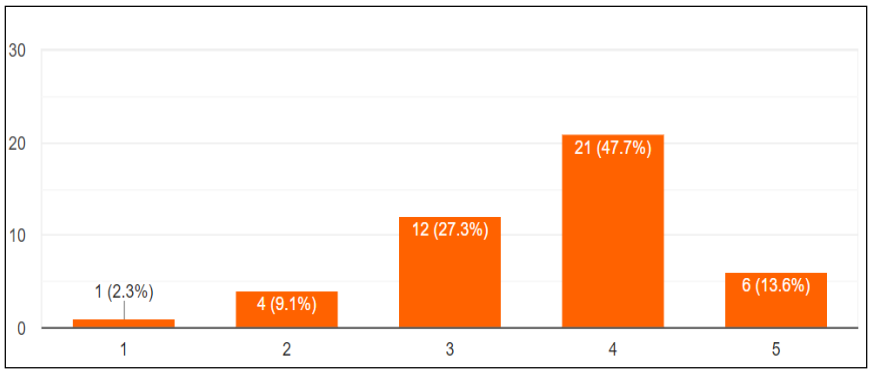

Figure 1: Experts Rating of various test Forms of Test 1

(b) which is the most difficult and which is the easiest test form (although they produce equivalence certificate)

Surprisingly same test form was judged as the most difficult as well as most easy.

C. Evidence from Measurement Models

We tried to corroborate perception with item analysis findings. Students responses were analysed using "Psych" in R, whereas RUMM 2030 was used for Rasch analysis. Only relevant parts have been discussed here.

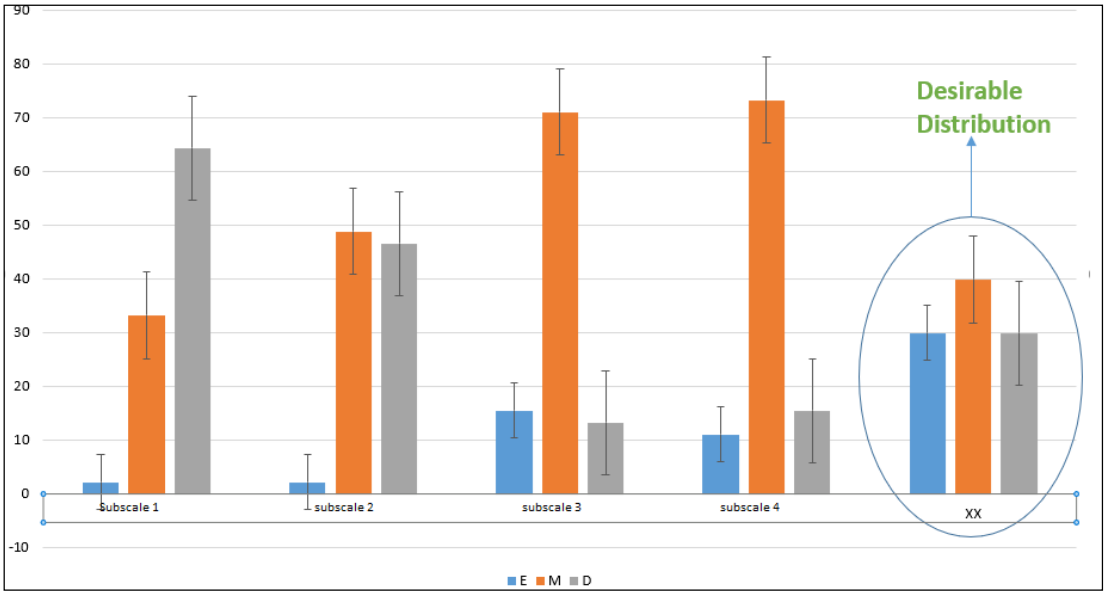

Figure 2: Distribution of item difficluty in Test Form A of Test 1

Figure 2 shows proportion of easy, medium and difficult test items of various subscales, based on item mean expressed in terms of average proportion of correct response by the test takers, as per CTT.

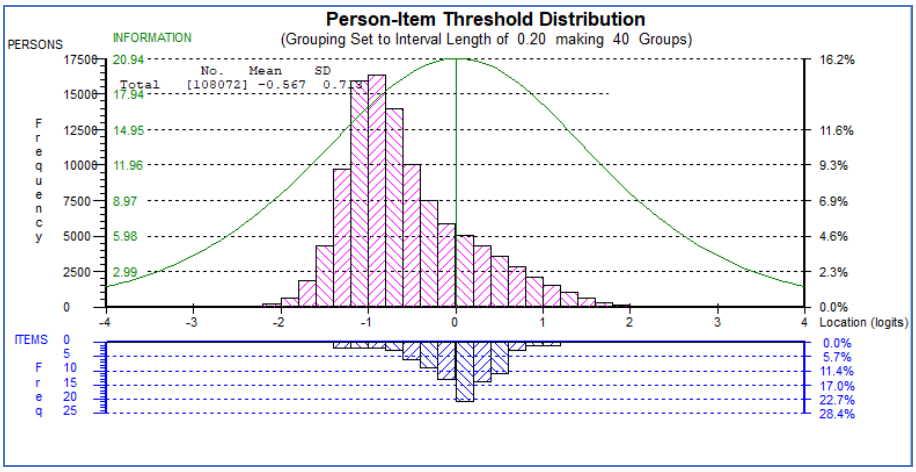

Figure 3: Test Targeting of Test Form A of Test 1

It can be inferred from Figure 3 that a major section of **candidates** is not targeted by the test, which has implications for policy makers, much beyond the technical aspects of educational testing.

To validate/refute authors' perception depicted in Figure 1, we tried to see perception in conjunction with dispersion measures.

Table 2: Distribution of candidates correct response on various test forms

Test forms	Quantiles			
	25%	50%	75%	100%
Form 1	12	17	23	85
Form 2	11	16	23	86
Form 3	12	17	23	84
Form 4	11	16	22	84
Form 5	12	17	23	84
Form 6	13	18	24	86
Form 7	12	17	23	87
Form 8	11	16	22	86

D. **Analysing evidence**

All of the above evidence was discussed with concerned group of subject experts to give a meaning to these qualitative and quantitative evidence, both at this holistic level and at the level of individual test items. Possible reasons were discussed, and strategies were redesigned. Thus, EBA can be visualized as a tool for building chain of reasons based on evidence, at every step of assessment. We have some indications of improvement in the preliminary analysis of next cycle, but we know that it is a long process.

Summary and conclusion

The study discussed here has been discussed as an example of a new beginning. In practice it is not so simple to deviate from routine practices. Especially, engaging experts and traditional test developers with years of experience of working in an accustomed manner is the most challenging task. Overcoming skepticism, unwillingness to work in a recommended manner, and doubts over modifying the practices that had served the developers well in the past is critical. Sometimes it appears as a conflict between subject expertise and assessment practices. However, after 1 cycle, majority of them show openness and a develop positive attitude towards practices. It is not a new practice different from good traditional test development practice. However, EBA process are specified in far greater detail, which enable test developers accomplish work more efficiently. However, flexibility and art of test development are as important in the practical application of EBA as they are in traditional test development.

Next area to leverage potential of EBA is eliciting valid responses from d**esign in order to address - How do we measure.** Effort is to identify appropriate model for adaptive testing, based on IRT measurement model and cognitive theories. Which algorithm would be suitable for which test, what should be batch or testlet size that would generate appropriate estimate of student's previous level, what is evidence that a particular batch size is sufficient to determine the next stage, two-stage or four-stage, etc questions demand valid and reasonable evidence to support a specific model.

References:

Gulliksen, H. (1961). Measurement of learning and mental abilities. Psychometrika, 26, 93–107.

Lewis, C. (1986). Test theory and Psychometrika: The past twenty-five years. Psychometrika, 51, 11–22.

MHRD. (1992). National Policy on Education, 1986 (As modified in 1992)" (PDF). Ministry of Human Resource Developemnt. India

Mislevy, R.J. (1994). Psychometrika. 59: 439. https://doi.org/10.1007/BF02294388

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). Evidence-centered assessment

design. Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design (Research Report 03-16). Princeton, NJ: Educational Testing Service.

NKC (2009). Report to Nation 2006-2009. National Knowledge Commission.

NRC (2001). The Nature of Assessment and Reasoning from Evidence." National Research Council. 2001. The National Academies Press. doi: 10.17226/10019.

Zeiky, M.J. (2014). An introduction to the use of evidence-centered design in test development

Michael J. Zieky. Educational Testing Service, Princeton, U.S.A.