Digitizing Specialized Assessments for Educational Professionals in the United Arab Emirates: A Conceptual Analysis and Innovative AI Driven Approach

Raona Williams, Ministry of Education, United Arab Emirates

The European Conference on Education 2023
Official Conference Proceedings

Abstract
This conceptual paper reveals novel insights on ways that digitized psychometric testing of subject matter and pedagogical competencies is developed as part of teaching professional credentialing procedures in the United Arab Emirates (UAE). As global fields continue to embed technological advancements, digitizing assessments is becoming more widespread in education. In the UAE educational professionals successfully complete digitized assessments as part of obtaining a license which assures alignment of high-quality teaching practices across the nation. Incorporating innovation, the use of evidence-based test design processes is bolstered by content developers, psychometricians, and international experts. The reader will understand how literature covering test theories, virtual professional learning community (vPLC) and community of practice frameworks bolstered in reflective practice underpins testing design. Supported by the author's specialist research expertise in networked learning, virtual professional learning communities of practice and educational theories, the paper uncovers how Rasch modelling analysis, collaborative situated learning and professional reflection models are used to steer committee experts and test specialist project managers as they contribute expertise in educational assessment, specialized subject matter knowledge, educational pedagogy and cultural awareness. There is also a concluding inference on how artificial intelligence and machine learning processes are used to facilitate and refine iterative cycles of test designs and evaluations with an innovative and efficient approach towards developing digitized ipsative assessment methods for education professionals working in the UAE. This paper is of importance in highlighting the UAE's robust approach of developing teacher licensure testing, with a unique perspective.

Keywords: Virtual Communities of Practice, Digitized Assessment, Item Response Theory, Professional Licensure

INTRODUCTION

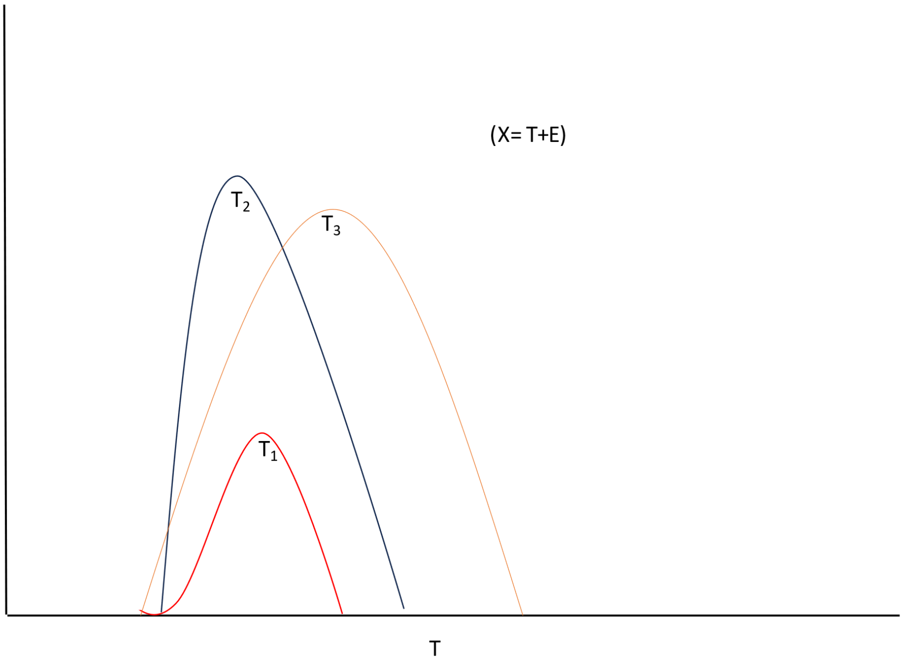
Maintaining a high quality of education underpins the building of a successful nation and measuring quality can be achieved by robust testing procedures. In line with future focused sustainable global goals, developing workforce competency components and bolstering further career professional development, monitoring and testing are important factors that guide success. Through recent decades, the face of testing has undergone changes. Measuring achievement in education is now increasingly being conducted through online assessments, administered and managed through the medium of computational technologies using electronic devices and the internet.

The United Arab Emirates' Centennial 2071 national vision (UAE, 2023) sets out aims towards achieving the best for its country through its future focused government objectives. Guided by cutting edge international research in educational assessments and digitization as technologies continue to move further forward into the web 3 era (Wan et al, 2023), there is a dedicated directorate in the UAE Ministry of Education responsible for the design, administration and evaluation of national and international tests set for different educational populations in the UAE. A diverse range of national tests are administered, some geared towards school learners known as the Emirates Standardized Test (EmSAT), a series of low stakes and high stakes tests which are designed to chart the progress of students through their school years and measure their college readiness for entry into the tertiary system of education. Other tests, known as Teacher Licensing Systems (TLS) specialization tests are developed and well-grounded on principles and empirical evidence in psychometric assessment practices and procedures. They are high stakes tests geared towards validating education professionals who are tasked with facilitating student progress within the nation. These tests cover domains of subject matter content in relation to candidates chosen teaching specialism. Irrespective of the educational discipline or position, whether it is for teachers, school leaders, or other school staff such as counsellors, special needs coordinators or technicians, the license indicates how UAE-based education professionals are equipped with desired competencies for providing excellence in theoretical and practical knowledge, providing records of the ability to compete globally. Therefore, these TLS high stakes tests which forms part of professional licensing processes within the UAE, are completed by education providers who have successfully completed preceding stages linked to attesting their qualifications and demonstrating proficiency in pedagogical teaching techniques and skills. The paper focuses on outlining evidence-based approaches incorporated in test design development and evaluation, including classical test and item response theory (Lord, 1980; Wright, 1968), virtual professionals learning communities (vPLC) (Hord, 1996) communities of practice (CoP) (Lave and Wenger, 1991) and Schon's (1983) reflective practice framework employed in practice.

DIGITIZING SUBJECT SPECIALIZATION ASSESSMENTS

Test Theory Frameworks: Classical Test and Item Response

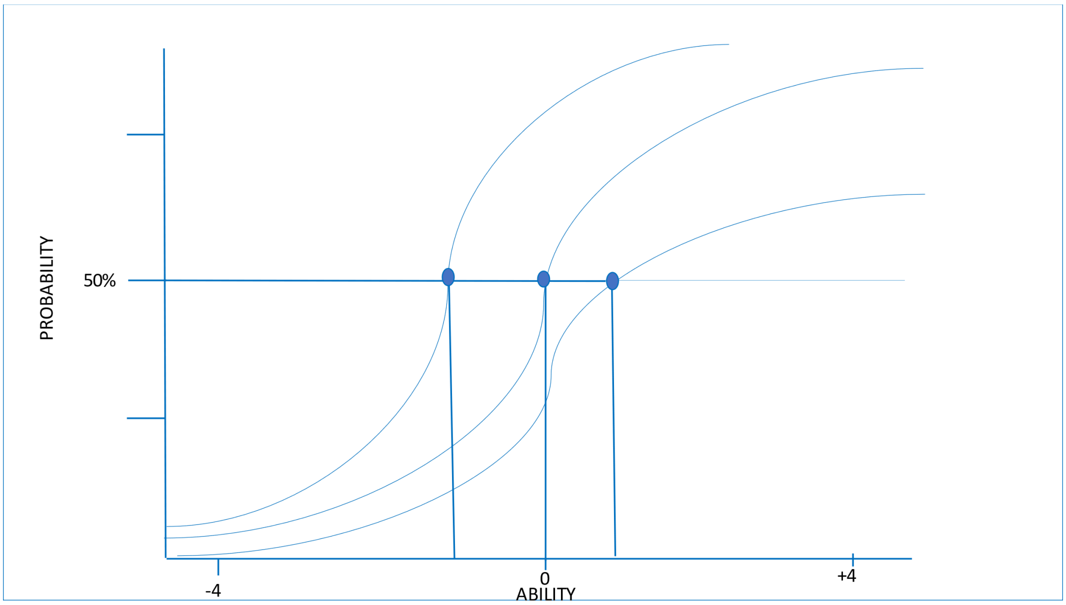
Classical test theory (CTT) and item response theory (IRT) frameworks are implemented in the design and scoring evaluations of psychometric assessments. CTT scores with a traditional sum-of-points approach whereas IRT scores with a latent scale approach. They provide a quantitative representation of test item quality and content validity. From the latter half the last century, test theory frameworks have been evolving and through the work of psychometrician and mathematician scholars such as Frederick Lord (1980) and George Rasch (1960), IRT is becoming more firmly fixed as part of the paradigm shift that is occurring in educational fields. CTT frameworks, which precede IRT frameworks, analyze tests based on a theory of a combination of true scores with an element of error. The error element is estimated using reliability coefficients and is useful when the questions to be administered are of a uniform design, that is, a design that does not consider question difficulty, discrimination, or a guessing factor. This theory carries an assumption of linearity between true score observations and error measurements, meaning that an observed score (X) is regarded as a true score (T) plus some degree of error (E). It also carries an assumption that the errors are normally distributed with a mean score of zero. This is typified in Fig 1:



(Fig 1: Classical Test Theory [X=T+E] assumptions)

It is important to indicate how well a test evaluation considers subject content alongside latent traits such as variation in candidate abilities. Literature denotes CTT as a traditional approach that has limitations when considering latent factors and this has resulted in IRT being more common practice in the development of digitized assessments using theories and techniques known as mathematical Rasch modelling (Linacre, 2002). Rasch modelling involves applying mathematical probability calculations to the evaluations of responses to an item bank of questions administered. The calculations consider candidate performance, measuring candidates' abilities in a personalized manner. Thus, compared to CTT, IRT is a more modern framework that incorporates latent traits. It serves to analyze test items with a statistical prediction applied to responses based on properties related to item difficulty and candidate ability. The predictions involve probability based mathematical calculations and algorithms that consider candidate ability and the likelihood of them achieving the correct answer based on the difficulty of the question. Probability estimates are comprised of three parameters, one measuring student's ability, one measuring item difficulty and a third measure indicating its discriminatory feature. Involving processes of machine learning from human behavior to improve problem solving, saving time, and reducing analytical efforts, it is now used in educational test design and making predictions of individual capacity. IRT theory has added to classical test theory, incorporating mathematical models to improve the testing process (Weiss and Kingsbury, 1984). It incorporates both accuracy and fairness because it incorporates the difficulties of test items alongside examinees abilities (de Ayala,

2009) thus providing an analysis of the relationship between test item characteristics, candidate ability and their test scores. Whilst IRT is a modern approach, it is important to note that test frameworks are chosen respectively to achieve set objectives in test design. Therefore classical test theory can still be an appropriate framework if say, tests are of the same difficulty. This is not the case for digitized assessments administered to educational professionals in the UAE. The primary testing theory used for these TLS assessments is the Rasch one-parameter logistical model (Rasch, 1960) which follows dichotomous scoring. A detailed treatise on how mathematical analysis is applied in scoring of test items will not be given in this paper, however as an overall summary, key indications of test items are given on item characteristic sigmoidal (S) curves (Figure 2), which indicate the degree/percentage of probability between observations of how a candidate performs against expectations of how a candidate should perform. Each item has characteristic curves based on their item difficulty. A nominal reference of ability (0) is assigned to an item relative to their difficulty that indicates a 50% probability of success with infinitesimal indications left or right of that reference.



(Fig 2: Item Response Theory characteristic curves)

With a critical limitation of IRT being that replication of psychometric estimates is dependent on sufficient sample sizes, items must be developed within an item bank and through iterative test administrations and evaluations, they are scored with a consideration of both the candidate ability and item difficulty with an assigned mathematical probability based on modelling and statistical functions. Fundamentally, it involves an odds ratio of probability representation – or a logarithmic function 'logit' representation. Using a logit function to represent relationships, IRT carries practical advantage in that the computation involved provides a powerful context to draw robust inferences.

Test items are developed by subject matter experts and become populated in an item bank. An item bank is a composition of carefully designed questions that develop, define, and quantify a common key indicator or theme, thus providing a variable that can be measured (Wright and Bell, 1984).

As opposed to an item considered on its own, an item bank houses an ever-expanding population of questions that undergo a series of calibrations following test administrations and has associated psychometric features in relation to its design quality. The quality and calibration of test items are informed and guided by scoring and evaluations from successive administrations. The performance of test items is measured and continually represented through the infinitesimal growth of the S-curve generated by mathematical probability calculations. Candidates that sit the test of a certain ability will achieve a specific score on a test item.

Educational assessments designed by global organizations around the world, driven by market forces, have gradually shifted from paper based to online assessments. In tandem with this, there have also been developments towards incorporating computational psychometrics with semantic technologies, artificial intelligence and machine learning applications. With advances in computational power over the last decade, storage with rapidly increasing volumes of data and greater efficiencies in sharing of virtual information, there are now assessment system software companies that specialize in aligning the theory and mathematical application of IRT digitized assessments. The UAE MOE uses the item calibration software Winsteps ® (Linacre, 2023) to evaluate and inform assessment cycles accordingly. In accordance with an increasing prevalence of item response theory test designs, well established software companies continue to develop platforms that provide testing specialists with viable options to include artificial intelligence (AI) methodology in their assessment development. As a result, this facilitates greater efficiencies for the UAE Ministry of Education directorate to lead in combining technology advancements, big data management, and algorithmic computational analysis as they develop national and international assessments.

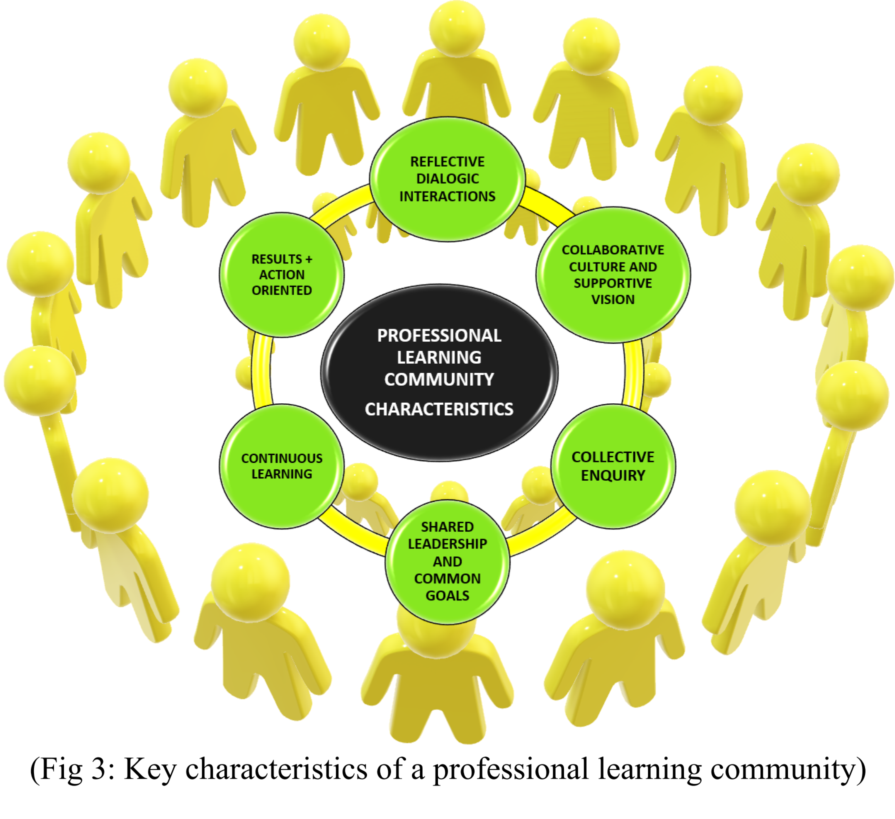
DIGITIZING SUBJECT SPECIALIZATION ASSESSMENTS

Virtual PLC/CoP Committees Incorporating Reflective Practice and Complementing AI Guided Test Development

With a population comprised of over 200 different nationalities from all over the world living and working alongside national citizens across the seven emirate regions in the UAE, it is a leading country that is recognized for its innovational growth, and an established region of the world that attracts international teaching professionals to work in public and private sector educational organizations. Variation in licensure processes to certify teachers exists across the world and this requires detailed understanding as professionals come from a diverse range of countries. Added to this, there are key UAE policy directives which encompass high standards and appreciations of unified cultural values across each emirate that should be exemplified in relevant areas when considering testing design. To represent this unique breadth of considerations, building future focused knowledge-based professional learning communities of practice is a central pillar in iterative project management procedures for test cycle development stages. Ipsative assessments are carried out in recurring cycles to inform item bank development with both quantitative evaluations and a subject matter expert's reflective and qualitative input. These high-quality learning spheres and professional communities convene with a hybrid approach, engaging in discourse through both in-person and virtual settings to support the burgeoning of digitization in the country across varying sectors. This is guided by framework models of professional learning communities postulated by Hord (1996) communities of practice developed by Lave and Wenger (1991) reflection models such as that developed by Donald Schon (1983).

Virtually mediated communication tools which facilitate professional learning communities of practice have grown in popularity through the Web 2 era (DiNucci, 1999) and are now commonplace in educational settings. There are various platforms centered around digital technologies and software applications which facilitate critical discourse have been implemented such as, videoconferencing, online blogs, discussion boards and social media messaging applications (Carlen and Jobring, 2005; Duncan-Howell, 2010; Williams, 2018). Learning management platform systems such as Microsoft Teams ® and environment hubs like Google Meet ® used to house virtual learning communities over recent years has become vital and can meet the multifactorial needs in enabling cyber secure big data capture and management by public and private sector organizations. Understandably, the implementation of virtual PLC's (vPLC) using these online settings have been popularized due to their suitability in facilitating smarter ways of working.

Cognitive anthropologist Jean Lave, and educational theorist Etienne Wenger are seminal pioneers of Communities of Practice (CoP) a framework model to maximize purposeful collaboration and problem solving (Lave and Wenger, 1991). Additionally professional learning community models postulated by Hord (1996), places reflective dialogue as a central focus with success being achieved through its typified framework characteristics shown in Figure 3. These characteristics underpin committee member focused actions. With collective enquiry, committee members engage in vPLC environments with a concerted focus on making continuous improvements as they work through reviewing and preparing test item content using reflective practices. Practical approaches to aid reflective practice can vary in dialogical communication modes using visual, audio, and written formats. Virtual spaces that capture information in real time, encourages communicative engagement, creative participation and deep situated learning (Selwyn, 2016). From an educational perspective, considering the construct representation of pedagogy, subject specialist domain areas and statistical data elements in test item development, reflective practice aids the quality of expertise shared.

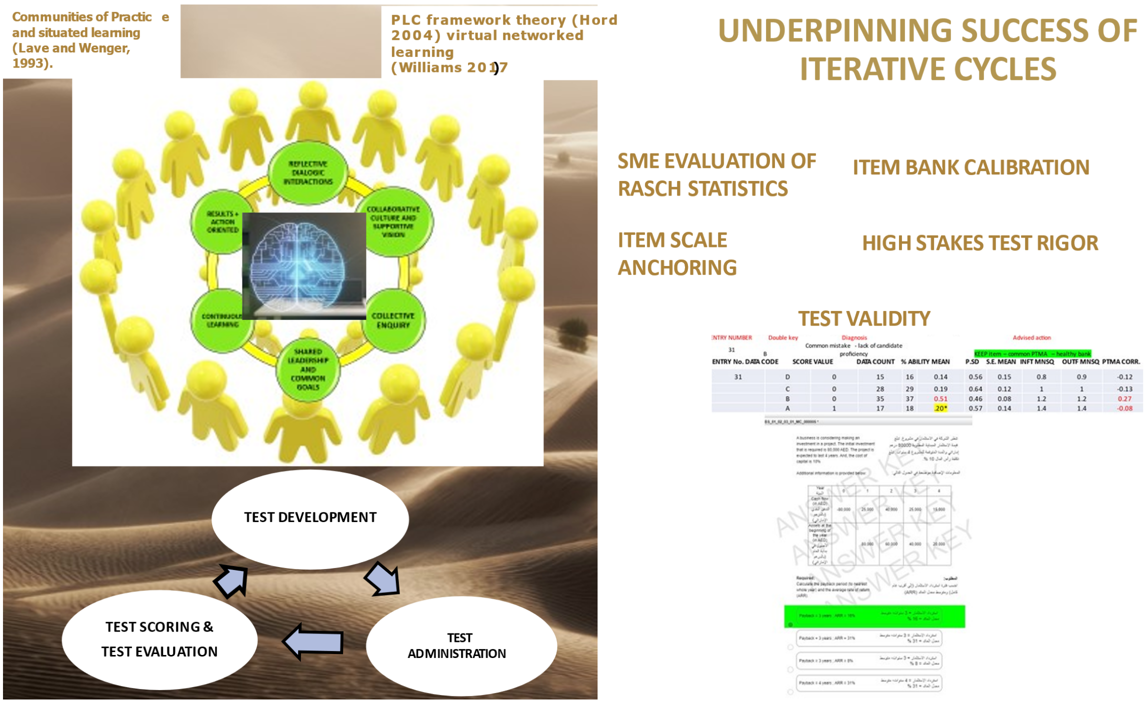

REFLECTIVE DIALOGIC INTERACTIONS
COLLABORATIVE CULTURE AND SUPPORTIVE VISION
RESULTS + ACTION ORIENTED
PROFESSIONAL LEARNING COMMUNITY CHARACTERISTICS
CONTINUOUS LEARNING
COLLECTIVE ENQUIRY
SHARED LEADERSHIP AND COMMON GOALS

(Fig 3: Key characteristics of a professional learning community)

In line with international standards in test development (APA, 2000), TLS digitized assessments are developed from an internationally benched framework which details subject specific standards, grounded by a critical review from a range of continents such as the Gulf region, USA, UK, Europe and Australia. Test frameworks are developed and reviewed by expert stakeholders from a range of national and international organizations which make up the professional learning community before being ratified. Experts are comprised of educational and industrial professionals from the all over the world involved in curriculum development, governmental policy, higher education and subject content based career fields. There is also concerted focus on aligning test item development with international strategies, related specific subject matter areas and remaining in accordance with laws and practices relevant to the UAE constitutional decrees. Frameworks developed for each specialization test provide comprehensive detail on testing domain areas, providing rationale for topic inclusion, item difficulty, complexity and scaffolded competencies. Guided manuals for candidates are also developed and reviewed by committee members which lay out the test scope and indications of how tests will be administered. A selection of test items is given which provide information on the question format. vPLC committee members are then dedicated to the wide development of customized questions which are populated within a computer-based test item bank. This item bank contains a range of questions that are prepared by subject matter experts in domain areas, reviewed both internally within the directorate and externally by experts from around the world. Each item is also carefully designed to enable clear duality in languages English and Arabic – this is unique.

The vPLC settings complement the iterative nature of project management involved in digitized assessment development, it is a vital resource to facilitate situated learning amongst committee members. Clear records of communication can be shared to all, real time patterns and responses to situations are readily available and then used to highlight areas for committee members to be aware of when considering inherent criticalities, thoughts and actions which guides their semantic development of item constructions and reviews. Whilst the subject matter is evidence based and grounded through international benchmarked frameworks, it is important to consider test rigor and validations for these high-stakes examinations. Factors to address include justifying concepts around the semantics employed in the test construct, the plausibility and quality of test item distractors and associated translation concerns with the dual language nature of the test item development being in both Arabic and English. The quantitative data collected for test items can be analyzed in depth with bespoke vPLC members taking into consideration their subject matter and industry expertise with an underpinning critical eye to ensure cultural and national visions and practices are upheld.

Discussing quantitative statistical parameters which provides an indication of test quality can be analyzed in multimodal fashion, and then corroborated or disputed when liaising with subject matter experts and the international review committee members relevant to each subject. Through a systematic review of quantitative psychometric data with a subject matter expert analysis of test item quality, certain items are deemed to be exemplars and serve as referenced material to improve and further guide the validity and accuracy of the tests as their growth and development continues in successive cycles. In concordance with CoP and PLC framework models, an authentic voice from members can be captured through both observed interactions and vPLC discourse guides test item formations from a shared 'reflection-in-action' problem solving perspective. By gathering information from focus group discussions and individual committee member responses, interactions and reflexive interpretations, qualitative data collected can then be re-referred to with a deeper reflection-on-action model approach (Schon, 1981) at later stages of the iterative cycle. This is notably advantageous to facilitate the steering of the community by content developers, psychometricians and project managers when making evaluative decisions about quantitative statistical data pertaining to item calibrations. Through dedicated analysis of quantitative statistical information alongside qualitative rich discussions taking place in the growth of item bank populations, it can therefore be surmised that this fresh and unique approach creates a human centered subject matter expert (SME) 'prompt engineer' styled outlook using artificially intelligent data to drive developments in bolstering content creation and validity of digitized assessments. Thus, offering a unique perspective on how a global vPLC contributes an innovative approach towards digitized assessment development for UAE international education professionals.


Communities of Practic e and situated learning (Lave and Wenger, 1993).
PLC framework theory (Hord 2004) virtual networked learning (Williams 2017
REFLECTIVE DIALOGIC INTERACTIONS
COLLABORATIVE CULTURE AND SUPPORTIVE VISION
COLLECTIVE ENQUIRY
SHARED LEADERSHIP AND COMMON GOALS
CONTINUOUS LEARNING
RESULTS + ACTION ORIENTED
TEST DEVELOPMENT
TEST ADMINISTRATION
TEST SCORING & TEST EVALUATION
UNDERPINNING SUCCESS OF ITERATIVE CYCLES
SME EVALUATION OF RASCH STATISTICS
ITEM BANK CALIBRATION
ITEM SCALE ANCHORING
HIGH STAKES TEST RIGOR
TEST VALIDITY
ANSWER KEY

ENTRY NUMBER 31 | Double key B | Diagnosis: Common mistake - lack of candidate proficiency | Advised action: KEEP item – common PTMA – healthy bank

ENTRY No.	DATA CODE	SCORE VALUE	DATA COUNT	%	ABILITY MEAN	P.SD	S.E. MEAN	INFT MNSQ	OUTF MNSQ	PTMA CORR.
31	D	0	15	16	0.14	0.56	0.15	0.8	0.9	-0.12
	C	0	28	29	0.19	0.64	0.12	1	1	-0.13
	B	0	35	37	0.51	0.46	0.08	1.2	1.2	0.27
	A	1	17	18	.20*	0.57	0.14	1.4	1.4	-0.08

(Fig 4: Successive iterative cycle stages of UAE Digitised TLS Specialized assessment)

DIGITIZING SUBJECT SPECIALIZATION ASSESSMENTS

Further Research Considerations

As typified in figure 3, over the course of each stage of successive iterative cycles, test items are calibrated accordingly to serve as exemplars, or anchors in the mathematical model evaluations of further test cycles and informing committee member item writers and reviewers. Regular discussions are held with vPLC committee members to discuss test item quality based on the statistical results given, and there are times that items may then have to be redeveloped, removed or identified accordingly based on qualitative and quantitative analysis. Incorporating national and international specialists who are experts in their academic and industrial fields, project managers within the directorate steer this powerful brain center for the purposes of growing the item bank accordingly with calibrated and well anchored items. Whilst procedures such as standard setting and scale anchoring of items is incorporated into connecting test content evaluations to scored interpretations, further development is being incorporated to consider differential factors that affect analysis in relation to the candidates. (Cowan et al, 2020). As with the nature of teacher licensure assessments, variations in sample size of candidates, their performance rating and their experiential levels are areas to consider for further research.

The fast-paced developments occurring within the web 3 era create challenges. Whilst there are benefits to keeping abreast of latest technological trends, there are controversies surrounding the ethical sustainability of areas such as generative artificial intelligence and associated machine learning that seem to be developing an overarching presence in the world today. Therefore, it goes without saying that moving towards a digitized system of assessment carries with it associated concerns with cybersecurity. Areas related to securing browsers, embedding remote proctoring, encrypting data, authentication of Ips and integrating AI based deep learning strategies are most certainly considerations that further research can be centered on.

DIGITIZING SUBJECT SPECIALIZATION ASSESSMENTS

Conclusion

This paper has provided conceptual critique revealing how digitized assessments are conducted for educational professionals working in the gulf region. It has provided insight on evidence based literature and frameworks which underpin assessment design and administration of specialized digital assessment for international education professionals working in the Gulf region A treatise of classical testing theory and item response testing theories was given, with contemporary considerations on virtual networked learning communities and reflective practice frameworks. The paper has uncovered a novel way to use these frameworks to capture subject matter experts' contributions in an open, transparent way and through a shared AI driven critical lens; inform, engineer and bolster test design procedures. This creates further ideas towards strategies that can be implemented incorporating human expertise, varying depths of and situated learning within a global community of academic experts to facilitate the building of content validity and test design mastery over consistent and iterative cycle stages when preparing digitized tests for education professionals.

The UAE Ministry of Education is embedding the digitization of their assessments into their testing strategies and division developments. This is to ensure the country is ready for the future in line with their Centennial 2071 vision. They want their residents, citizens and those that are tasked with the responsibility of delivering excellent education to their society to be adaptable, flexible and future focused. As market forces continue to drive web-based technologies, new ways of working, designing, and developing the quality of measuring education performance across the sector with an aspirational and pro-active perspective is being adopted by UAE project managers, content specialists and psychometricians within the national testing directorate. Conscious processes such as these which enable the capture and analysis of quantitative and qualitative big data, bolsters the validity, accuracy and rigor of these high stakes digitized assessments. Therefore, laying a concrete foundation for further technologic developments involving machine learning techniques and generative AI technologies.

REFERENCES

APA (2000). Standards for Educational and Psychological Testing. Washington DC: American Educational Research Association.

Carlen, U., Jobring, O., (2005). The Rationale of Online Learning Communities. International Journal of Web Based Communities 1(3), 272-295.

Cowan J., Goldhaber D., Jin, Z., Theobald R., (2016). Teacher Licensure Tests: Barrier or Predictive Tool? CALDER Working paper No 245-1020.

De Ayala, R.J. (2009). The theory and practice of item response theory. New York: The Guildford Press.

DiNucci, D., (1999). Fragmented future. Print Magazine, 4 (32).

Duncan-Howell, J. (2010). Teachers making connections: Online communities as a source of professional learning. British Journal of Educational Technology, 41(2), 324-340.

Hord, S., (1996). Professional Learning Communities: Communities of Continuous Inquiry and Improvement. Southwest Educational Development Laboratory.

Lave, J., and Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge University Press.

Linacre, M. (2002). What do Infit and Outfit, mean square and standardization mean? Rasch Measurement Transactions, 16, 878.

Linacre, M. (2023). Winsteps ® Measurement, Version 5.5.0 [Computer Software] Winsteps.com.

Lord, F.M, (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rasch, G. (1960). Probabilistic model for some intelligence and achievement tests. Copenhagen: Danish Institute for Educational Research.

Selwyn, N. (2016). Social media and education. Now the dust has settled. Learning Media and Technology 41(1), 1-5.

Schön, D. (1983). The Reflective Practitioner: How Professional Think In Action. New York.

UAE (2023c). UAE Centennial 2071. UAE centennial plan. Available online at: https://uaecabinet.ae/en/uae-centennial-plan-2071 (accessed 10 August 2023).

Wan, S., Lin, H., Gan W., Chen, J., Y P.S. (2023). Web3: The Next Internet Revolution. Available online at: https://arxiv.org/abs/2304.06111 (accessed 10th August 2023).

Weiss, D.J. and Kingsbury, G.G, (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21(4), 361-375.

Williams, R., (2018). What's App with the Social Message? Networked Learning and Social Media Interactions in Curriculum Design: Exploring Student Perceptions. Caribbean Journal of Education 40 (1&2) 220-236.

Wright, B.D. (1968). Sample-free test calibration and person measurement. In Proceedings of the 1967 invitational conference on testing problems (pp85-101). Princeton, New Jersey: Educational Testing Service.

Wright, B.D. and Bell, S.R., (1984). Item banks: What, why and how. Journal of Educational Measurement. 21 (4) 331-345.