Asia-Pacific Journal of Educational Management Research
Vol.4, No.3 (2019), pp.21-32
http://dx.doi.org/10.21742/ajemr.2019.4.3.03

Learning Evaluation Methods in University based on Data Mining

Golshah Abawajy[1]

Charles Sturt University, Wagga Wagga, Australia
Abawajy5749@gmail.com

Abstract

One of the current frontier points of learning evaluation is to focus on whether and how to use powerful digital technology to analyze digital data. This paper proposes a learning evaluation method based on big data. This paper constructs a new standard for the development of evaluation tools—metrolytic standards. The combination of standards used in the field of learning analytics and commonly used methods in educational measurement provides a framework for ensuring the reliability and validity of all educational evaluations. Measurement and analysis standards include quality requirements for the reliability, validity, accuracy, or interpretability of the test. These requirements are usually only applicable to high-risk, large- scale evaluations, such as PISA, SAT, or GMAT. The application of measurement analysis standards is based on a solid understanding of evaluation and its role in learning, combined with the advantages of learning analysis, artificial intelligence, and measurement science, and provides a choice for researchers in the frontier field of evaluation.

Keywords: *Digital technology, Artificial intelligence, Learning evaluation tools, Learning evaluation methods*

1. Introduction

It is not uncommon to think about how to change the method of learning evaluation. All along, the practice of learning evaluation has closely followed the focus of education and society [1]. For example, contemporary psychometric methods can be traced back to the 19th-century anthropologists' and eugenicists' interest in individual differences. At that time, Ronald Aylmer Fisher, Charles Edward Spearman, Karl Pearson, and other statisticians developed a series of methods to identify individual characteristics, many of which are still in use today. The multiple-choice test questions that appeared in the early 20th century aimed to objectively and fairly evaluate applicants for positions in the US military on a large scale and provide a reliable ranking. In the 1980s and 1990s, as schools needed fairer selection methods, and at the same time needed to take greater responsibility for students' learning, standardized testing methods were improved in this context. The combination of sophisticated statistical methods and automated measurement technology enables large-scale evaluation and monitoring to be achieved, which is often widely used to assess "scholastic aptitude" or students' mastery of basic learning content such as literacy and arithmetic.

Nowadays, the digital tools used in university teaching are changing the way of learning evaluation and the relationship between evaluators and evaluators. Teachers use digital tools to

Article history:
Received (July 2, 2019), Review Result (September 8, 2019), Accepted (November 16, 2019)

Print ISSN: 2207-5380, eISSN: 2207-290X AJEMR

set up an evaluation system to support student responses, supervise cheating, collect and score student responses, provide feedback and score to students, compile student response data, and form reports. The digitization of the learning evaluation program is more efficient in implementation, can be extended to large-scale classrooms, can reflect individual-level performance in a more targeted manner, is more interactive, and supports more imaginative, colorful, and interactive timely feedback, and can generate evaluation reports faster and more directly. Digital evaluation makes evaluation methods (such as peer evaluation and self-evaluation) that can only be operated with the support of complex implementation processes more feasible. Embedding the learning evaluation into the digital learning management system can better realize the teaching and learning, and provide support for the development of formative evaluation. Nowadays, digital media for teaching and evaluation through widely used learning platforms are very common, even in small classrooms on campus. It can be said that these technological advancements have made the process of learning evaluation more efficient, quicker, more sensitive, more formative, and timelier.

However, in addition to technical improvements in teacher evaluation learning, new frontier areas of evaluation are gradually emerging. Through exploring the use of powerful digital technology and digital data, especially Process data, to better evaluate and report learning results (especially the improvement of complex abilities and general knowledge). This article focuses on exploring this frontier field and analyzing the difficulties in developing effective and reliable learning evaluation. This article believes that it is necessary to build a new standard for the development of evaluation tools—Metrolytic standards. The combination of standards used in the field of learning analytics and commonly used methods in educational measurement provides a framework for ensuring the reliability and validity of all educational evaluations.

2. Contemporary pressure on learning evaluation

Contemporary universities are under pressure to change the content and methods of students' learning, and teachers' evaluations and reports on students need to be changed accordingly. In this context, the frontier of learning evaluation came into being[2]. Its core concept is that the "Fourth Industrial Revolution" is underway. Unlike the requirements of previous generations, the current era requires educators to cultivate learners with different technical capabilities [3][4][5], instead of following a single、Simple learning orientation. With the popularization of digital communication and computer technology, the rapid expansion of knowledge, and the impact of globalization, in line with the firm commitment to the sustainable and fair development of human welfare, the way of life and work in the 21st century is being redefined. The net effect of these series of factors on the school is: the school must redefine the scope of knowledge that students should learn so that students have other characteristics besides mastering the relevant knowledge in a specific field. In other words, students not only need to be proficient in content mastery but also need to master Know-how in a domain of study. At the same time, in addition to the training of cognitive ability in traditional disciplines, curriculum reform also requires learners to develop common knowledge, values, attitudes, skills, and beliefs in various disciplines[6]. This is the transition from knowledge to ability in course learning[7]. Therefore, course learning not only refers to the knowledge of the subject or professional field but also covers "Soft" or "21Century skills 21", "Transversal skills" and the "General Capabilities" described in this article, applicable to any field or category [8][10]. In 2015, the “World Economic Forum” listed the general skills that learners need to master, including critical thinking, communication, creativity, cooperation, scientific literacy, information and communication technology level, perseverance and curiosity, etc. [10].

Therefore, students need to develop lifelong learning skills[11]. This means that it is not enough for students to learn under the guidance of teachers in a formal educational environment. They must be able to learn independently. Modern learning ability is no longer equivalent to IQ or talent, but more refers to mastering a set of knowledge, skills, understanding, and belief in learning, to more or less enable individuals to have the learning literacy they need[12].

It is a professional challenge for teachers to formally incorporate general knowledge into the curriculum. For example, for courses that use traditional higher education evaluation methods, at the end of the course, teachers can make a summary evaluation based on the course essays submitted by students. Now, the key to the challenge is how to evaluate the learner's mastery of general knowledge.

It is a new field for teachers. Especially in large-class teaching, teachers do not necessarily know the situation of students, so it is more challenging to evaluate students' abilities. The assessment of students' complex abilities is very complicated and often needs to be carried out in a non-standardized environment, such as hand-made, cooperation with peers or teamwork, etc. Mastering complex abilities usually requires time and practice, and unlike traditional classroom teaching, the cultivation of these abilities needs to be carried out in a "real" learning environment. In addition, in each stage of learning, teachers need to give feedback on students' performance, so that students and teaching assistants can plan together and help learners gradually accumulate corresponding abilities. Therefore, at the same time as the curriculum reform, the goals and methods of learning evaluation should also be adjusted as necessary. Learning evaluation should assist students and teachers in judging students' mastery of the complex abilities and general abilities required in a certain field, and this urgently requires cutting-edge exploration of evaluation methods in this field.

3. The application prospects of big data and artificial intelligence in learning analysis

Based on the research results of large-scale international research projects aiming at "exploring how to effectively evaluate general skills", some scholars emphasized that the best learning evaluation method is "embedded in the technology applied in the learning environment, coexisting with technology and capable of mutual transformation" [13]. They pointed out that embedded technology can automatically generate feedback, provide on-demand evaluation, and prevent or reduce the separation of learning evaluation and learning experience. Since participants' activities are always reflected in the log stream, participants do not need to spend extra time and effort to collect data, so response rates are not a problem. Using the digital trails of learner activities for learning evaluation can calculate scores in real-time, and greatly improve the timeliness of calculation and feedback in the course of progress. On the surface, this method seems feasible. Nowadays, there is a large amount of learning-related digital information available, including clickstream data, which captures every mouse click, slide, or keyboard operation of all learners when using digital learning applications. Other information can be obtained through sophisticated digital data sensors in the classroom. These sensors can capture all information from the direction of the eyes to the heartbeat frequency, from speech to body movements. Therefore, the ability to "observe" what students say, do, do or write in the learning environment is greatly improved. The input information needed for traditional university learning evaluation includes the observations of teachers in the classroom, students' responses to evaluation tasks, or students' standardized test scores, but today's data

can systematically master all the information in the learning process, not just about Information produced by learning.

In addition, a large number of modern analysis methods can be used to analyze process data, such as social network analysis, text analysis, and various forms of data mining. The statistical data constructed by these methods can theoretically be used as a measure of student performance. Teachers can "observe" the degree of interaction between students and their peers in class, the focus of interest, and the systematic nature of learning habits through network analysis, text analysis, and time series analysis. Process data and corresponding analysis are usually presented on digital dashboards or fed back to teachers and students in other forms[14]. Artificial intelligence can also be used to analyze these process data. Artificial intelligence refers to the ability to simulate human intelligence through computer system programs[15]. In the past 10-20 years, artificial intelligence has been increasingly used in education evaluation. The emergence of artificial intelligence provides an opportunity for the development of more effective measurement tools. More effective measurement tools can objectively, effectively, and efficiently measure some traditional measurement methods and characteristics that are difficult to evaluate with data. At the same time, artificial intelligence can also help the development of new tests, and evaluate the development of learners' high-level skills (for example, critical thinking, cooperation, communication, and learning ability in an online environment) in the 21st century, and can make these evaluations more important [16]. The evaluation of these high-level skills relies on richer data, especially process data. In this digital age, these data can be collected through different channels. And artificial intelligence technology is conducive to the analysis and mining of these data, thereby forming an evaluation of students' high-level skills. The use of artificial intelligence technology to analyze MOOCs data to evaluate students' participation in MOOCs is an example. Sandra and her collaborators used machine learning technology, combined with educational and psychometric methods to analyze the discussion topics posted by students in the MOOC system forum [17]. Using artificial intelligence technology, they developed a method to automatically analyze the discussion topics posted in the MOOC system forum. However, this kind of analysis, if done by manpower, is very time-consuming and almost infeasible. The application of the topic model method based on artificial intelligence can automatically discover topics from unstructured data, analyze the frequency of the topics, and then convert these topics into indicators or topics. Modern psychological measurement models are used to analyze these indicators or topics to evaluate students' participation in MOOC learning, to predict their performance in MOOC learning. Artificial intelligence technology can also develop behavior indicators by analyzing the behavior and chat data of problem solvers in the process of cooperative problem solving to measure individual cooperative problem-solving capabilities [18]. Cooperative problem-solving ability is regarded as one of the core skills of the 21st century, which has attracted the attention of more and more researchers, educators, and employers. More and more researchers are trying to develop online tasks to record the process of problem solvers working together to solve problems. These process data include all the behaviors and chat records of the problem solver in the process of cooperating to solve the problem, and all records are time stamped. Researchers can use these behaviors and chat content to construct corresponding indicators, and then evaluate their ability to cooperate in solving problems. Artificial intelligence technology can help researchers automatically analyze chat content and corresponding problem scenarios, coupled with the analysis of behavioral data, can develop more effective and explanatory indicators, to more effectively measure the ability of cooperative problem-solving. All in all, the use of artificial intelligence to conduct deeper,

thorough, and efficient analysis of big data, especially process data, has broad prospects and provides a new paradigm for the future development of learning evaluation methods.

Researchers who entered the field of Learning Analytics early expect that process data will bring many benefits to students and teachers: realize the visualization of the teaching process, support teachers and students to reflect on teaching and learning practice; predict and simulate learning progress to achieve more effective learning interventions. Track and analyze each learner in real-time to realize the personalization of their learning [19][20][21][22]. If it is backed by artificial intelligence tools, then digital responses will be more capable of learning evaluation than humans.

The researchers' optimism is based on their belief that big data and artificial intelligence technologies are not only possible but should be the better choice when assessing learners' general abilities[23]. Traditional techniques for evaluating individual traits or abilities include the use of self-report scales, direct observation by experts, vocal thinking reports, analysis of subjects' diaries and other materials, and microanalysis methods (such as coding eye expressions or facial micro-expressions to infer individual behaviours) Traits), etc. [24]. But in a real learning environment, these technologies are impractical. They are costly and labour-intensive, so teachers and evaluators need to find better and more practical methods. Then, they will consider using digital big data provided by sensors embedded in the learning environment, which can systematically reflect all the information of learners in the learning process.

4. Ensure the validity and reliability of learning evaluation

Although the prospect of learning evaluation is optimistic, the existing difficulties cannot be ignored. Researchers in the field of learning analytics have always emphasized that digital big data, as a derivative of learning, is not necessarily better data [25][26]. Whether the digital traces of big data can be used to construct learning indicators, or whether they can effectively reflect learning results, this key question has not yet been convincingly answered. In addition, it is not yet known whether the process data contains sufficient information, and perhaps the missing information is precisely the necessary factor that can explain learning. Platforms or digital sensors cannot capture all "offline" activities, such as reflection, note-taking, or students' thinking activities, but this missing information may be crucial [27]. Researchers usually use correlation analysis, factor analysis, cluster analysis, and other methods to explore interesting rules based on large natural data sets. If the laws found are consistent with common-sense judgments and statistically significant, these laws have explanatory value and can be explored for their significance in learning. However, researchers do not understand whether these interesting and statistically significant laws are applicable to judge individual learning. They may just happen by accident, or they are of little importance to learning or even have no explanatory value. The statistical relationship can only show that the relationship is not random, but it is not enough to explain the results of individual learning measurement.

When using process data to evaluate and report the degree of improvement of complex abilities and general abilities, the most critical thing is the combination of analytical methods and educational measurement methods [28][29][30][31][32][33]. Measurement principles and techniques based on Mark Wilson's Constructing Measures Approach or evidence-based Evidence Centered Design Approach strengthen the credibility of the evaluation by ensuring that the value measurement of scores meets the necessary standards. When using scores to evaluate individuals, it is necessary to go through a careful, methodical, and course-centric

measurement process, which includes Development of Constructs and Evidence Maps, the use of specific rules and procedures to select evidence, and so on. The reason why measurement science establishes a standard is to use the standard to judge whether the measurement method is suitable for learning evaluation, that is, the selected measurement method should be effective and reliable, and can be used to accurately judge the individual's learning progress.

It is worth noting that when applying traditional measurement techniques to analysis-based process data, researchers are often very cautious. This reflects that scholars in the field of learning evaluation and analysis are increasingly aware that this type of data analysis is still in place. Initial stage. The most prominent difficulty in the frontier field of learning evaluation is that when using digital data to construct a measurement of complex capabilities, it is necessary to clarify the key assumptions that affect the quality of learning evaluation and test them one by one. For example, learning evaluation is always based on an assumption: Regarding a certain measured attribute, different individuals possess the attribute to different degrees, and the descriptive analysis of this attribute is the basis of evaluation. The trait itself must be meaningful and reasonable, and it has practical utility to evaluate it. This trait must have dimensions. People can understand why the trait differs from person to person. When evaluating the "more" or "less" of the trait, it must be possible to use the equivalent unit to measure all individuals consistently, and the equivalent unit can be accumulated and repeated. Even if the trait cannot be directly observed, observable behaviour differences (such as individual words and actions) can be used to explain the magnitude difference of the trait [38-39]. Individual differences in behaviour must have explanatory value, and it should be possible to infer the degree of this trait through these observable differences in behaviour. In short, when evaluating for education, such hypotheses should be tested one by one, to provide a basis for the relevant personnel involved in learning evaluation. The applicable standards for learning evaluation should refer to the discussion about validity in measurement science and the discussion about analysis quality in learning analytics.

Table 1 lists a set of "Indicative Standards" that can be used to examine the quality of the scores before all scores become the individual's final evaluation score. These standards come from measurement science and learning analysis practices, which are called "Metrolytics Standards" in this article. The word "Metrolytics" is derived from the Greek "Metron" (the root of measurement, meaning limited proportion) and "Analutikós" (meaning analysis). In an ideal state, measurement and analysis standards can provide a basis for the validity and reliability of the evaluation based on process data by the designer of learning evaluation, which is the same as the relevant high-risk test developers must provide a basis for proving the validity and reliability of their tests.

Table 1. Indicative measurement and analysis standards based on digitized process data

Standard	Explanation of the standard
Utility	When carrying out a learning evaluation, the goals must be clear, and the learning evaluation must have practical value to the relevant personnel.
Traits have clear characteristics	The characteristics to be evaluated need to have clear definitions. These characteristics can be expressed as different levels of mastery of knowledge, understanding, skills, beliefs, attitudes, or values. Stakeholders such as teachers and students should understand and accept the corresponding definition.
Traits have different dimensions	Suppose that different individual have different degrees of a certain trait, and this degree is continuous and measurable. In an ideal state, the measured degree can reflect the typical Learning Trajectories that which the learner gradually acquires the trait. These learning trajectories are also called Learning Continua or Progressions. Teachers and students should understand these learning trajectories.
Data is related to learning behavior	The selected data covers what learners say, do, do, and write during learning, and do not include personal characteristics that may be related to learning but do not affect learning outcomes (such as personal talents, socioeconomic background, or demographic characteristics, etc.).
Process data is "clean" and understandable	The methods of managing data include: checking the credibility of the data range and distribution and converting the original log stream data into numerical variables or categorical variables. Identify and minimize damaged, incomplete, misleading, or incorrectly compiled data; data definitions should be consistent and not change over time. Ensure that the time interval of data analysis is consistent with the purpose of data processing (for example, in time series research, the analysis unit should be accurate to seconds, hours, weeks, or years). Use analysis and sampling techniques to manage large volumes of data, etc.
Use statistical indicators to map data to learning progress	The data used to construct learning evaluation should be able to generate robust behavior indicators for each learner. For example, when using web analytics, the interaction between students may generate a statistical indicator of Connectedness. The differences in the measurement characteristics of indicators should be able to reasonably explain the differences in individual behaviors.
Consistent interpretation of indicators	Ensure that the evaluation scores of learning at different times are directly comparable, which is particularly important when using machine learning to construct evaluation scores or indicators. Once the algorithm changes the indicator, the structure under test may change. Any changes in teaching policies may also change the results of data inference. For example, under normal circumstances, students' voluntary participation in discussions can reflect their participation, but if they are forced to participate in discussions, their behavior reflects their obedience.
Indicators can fully reflect the characteristics	Behavioral indicators including scores can fully reflect different levels of this characteristic; statistical indicators will not be biased due to missing data or irrelevant data; there are no missing important features. For example, if offline learning activities are critical to the learning process, it is difficult to determine whether the online automated evaluation is biased.
Transparency of scoring and data conversion	A transparent index is established for the conversion from data to indicators to scores at each stage. At the same time, the entire measurement standards and algorithms should be clear at a glance.
Fully guarantee technical quality	The accuracy, discrimination, and reliability of statistical indicators and scores reflect the quality of psychological measurement: statistical indicators can reflect the completeness of learning and development, and the measurement scales are evenly spaced, and there is no obvious bias in any subgroup. These features can be tested by the fit of the measurement model.
Explanation of scores	There is only one reasonable explanation for the score, that is, it reflects the difference in the individual's ability level.
Discuss alternative methods	There is no more concise alternative method of learning evaluation.
Analyze the unexpected	There are no unexpected negative effects due to the shortcomings of the evaluation method.
Proposal for the possibility of review	If necessary, review the learning evaluation. Especially when using complex algorithms that are difficult to understand by the evaluation stakeholders for learning evaluation, the process of requesting a review is very important.

In recent years, the field of learning analysis has paid much attention to the rapid development of analysis applications and has begun to think about whether the analysis can provide a sufficiently credible evidence base. At the same time, considering that the results of learning evaluation usually affect the intervention that students receive next, the researchers expressed concern about this. Groups such as students, teachers, schools, or professional associations also have reasons to question whether measurement tools can evaluate complex abilities, especially when combined with different types of data and supplemented by complex data transformations or algorithms.

Therefore, the use of measurement and analysis standards has become a way to solve the above-mentioned problems. For learning evaluation that has both predetermined goals and practical value, measurement and analysis standards provide an evidence framework for the credibility of the evaluation. It should be noted that the measurement and analysis standards can not only support the evaluation results but also make people believe that the basic assumptions of the evaluation design or method can be verified. At the same time, there are no other reasonable alternative explanations for the evaluation content. Of course, this needs to consider all possible uncertain evidence and deterministic evidence and verify them one by one to reduce possible errors in learning evaluation.

5. The method challenge of using process data evaluation

The measurement and analysis standards listed in Table 1 point out the practical difficulties that analysts face when using process data to reliably and effectively evaluate complex abilities and general abilities. For example, measurement and analysis standards require that the content to be evaluated is clear and clear. In traditional classrooms, the evaluated content usually refers to the "teaching content" at the operational level. However, the evaluation of newer general skills requires clarification of the specific content to be evaluated. This requires researchers to clearly define the degree of knowledge, understanding, skills, beliefs, attitudes, values, and other characteristics of learners of different levels based on the understanding of learners' learning progress and trajectory. However, the difficulty often faced by designers of evaluation based on process data is that there are too few relevant cases describing the development trajectory of general knowledge. Therefore, when teachers or analysts design an evaluation method for a certain ability, the first task should be to define the learner's general learning process to reasonably describe the possible behaviour patterns of individuals who have mastered the trait to varying degrees. The learner's characteristics are measured on a potential continuous scale. Defining the learning process itself is not easy, and analysts or teachers often skip to the data and ignore this link. However, without a theoretical learning process defined based on empirical evidence, it is difficult to determine the validity, utility, and explanatory power of the evaluation score.

There are also many difficulties in evaluating individual performance, especially when evaluating general skills such as teamwork and collaboration skills that only exist in a social environment. Teachers are often well aware of the difficulties in evaluating such abilities. There is a complex relationship between the overall performance of the team and the performance of team members, and the data collected from digital forum participation, team collaboration, or multi-user interaction activities inherently possess such complex characteristics. The current measurement and evaluation methods are difficult to carry out a sufficiently reliable individual assessment based on such mixed data. It is worth noting that a recent report for the "National Assessment of Educational Progress" (NAEP) reviewed the relevant research on how to measure students' collaborative problem-solving ability based on

large-scale psychological measurements in the past 10 years. The report pointed out that, so far, we have not been able to reliably measure the ability of students to collaboratively solve problems, and how to solve this problem needs to be further explored. A non-technical solution is to focus the evaluation on the overall performance of the team rather than individual performance. Traditionally, the confounding effect of team attributes in evaluating individual performance is often regarded as "random error". However, some researchers who are more inclined to technology orientation believe that from the perspective of psychometrics, these different degrees of "errors" can in turn indicate the ability of the team after removing individual abilities.

In addition, it is necessary to further discuss whether the measurement of personal general knowledge such as problem-solving, interpersonal communication, or perseverance has universal significance, especially whether these abilities can be transferred to each other in different situations. For example, if a student shows good problem-solving skills in a chemistry class, does it mean that he can show the same outstanding problem-solving skills in a physics laboratory or other workplaces? The student cooperates with his peers in online games. It means that he can also cooperate with other members in face-to-face interaction. Early research shows that, with a few exceptions, these general skills are generally less transferable. Therefore, we need to be cautious about the evaluation results of complex comprehensive skills in specific situations, and we must fully recognize the limitations of this type of ability evaluation and its dependence on specific situations.

Teachers or learning evaluation designers may combine different sources of information when collecting evidence about complex abilities (for example, combining evaluations from different participants such as peers, self, and teachers, or combining evaluations from different forms of forums, lectures, etc.). However, if the relationship between various indicators from different information sources is not fully clarified, the quality of this comprehensive evaluation may be poor. Unless all indicators can reflect a certain potential ability in the same dimensions, the validity, reliability, accuracy, and practicality of the obtained evaluation results are very poor.

In the face of technical difficulties at the operational level, some scholars questioned whether we should not try to measure certain general skills. For example, G. Masters pointed out that "creativity" is a general ability that is notoriously difficult to measure in any environment, so we must first find out whether there is a universally meaningful ability of "creativity", at least in a specific field. Does it exist? Unless the relevant groups can agree on the definition and development process of a certain capability, the corresponding assessment is difficult to carry out.

It should also be pointed out that constructing a measurement that meets the measurement and analysis standards is not only time-consuming but also expensive and requires technical support. It is only economically feasible when used on a large scale, which in itself may limit the use of these methods in the short term.

The above discussion emphasized the practical difficulties in using process data to conduct high-quality assessments of general knowledge. The methodological challenges specifically include the lack of a clear development process of the relevant capabilities extracted from practice in the measurement process, the lack of universal significance of the evaluation results, the poor feasibility of the measurement standards, and the difference in the difference between groups and individual achievements. The technical methods used at the time have limitations and the integration of different sources of information needs to be treated with caution. Even robust measurement science methodology tools have a series of shortcomings. Therefore, we

must remain sceptical about any learning evaluation tool, and carefully consider its reliability, validity, accuracy, practicality, and interpretability.

6. Conclusion

In general, one of the frontiers focuses of contemporary learning evaluation is to focus on whether and how to use powerful digital technology to analyze digital data-especially process data-to better evaluate and report learning achievements (especially general skills). It should be noted that in the analysis process, the most likely error is to mistakenly equate "data" with "evaluation." The frontier of learning evaluation is extremely challenging, and the extent to which new big data sets—especially process data—can support effective and reliable learning evaluation is still unknown. Whether the current analysis method is suitable for the corresponding task, or whether the relevant personnel (including learners, teachers, and employers) trust the evaluation results are also uncertain. The use of process data is usually accompanied by the application of complex technologies, such as complex algorithms and data conversion. Therefore, it is inevitable that there are doubts about whether these data are useful for evaluation. In a digital, automated, and autonomous environment, inaccurate, unreliable, or even ineffective assessments of general knowledge will have a non-negligible impact on learners. Feedback and reports of learning evaluation will have a strong and real impact on learners, which may be positive, negative, or even destructive. Especially when the evaluation of intelligent learning will form the next learning intervention for students, if the evaluation result is contrary to the actual situation, the consequences will be worrying.

Measurement and analysis standards provide a reliable framework for learning and evaluating process data using learning analysis and artificial intelligence technology. However, the stringency of the standard may be too high to be implemented. Measurement and analysis standards include quality requirements for test reliability, validity, accuracy, or interpretability. These requirements are usually only applicable to high-risk and large-scale evaluations, such as PISA, SAT, or GMAT. The application of measurement analysis standards is based on a solid understanding of evaluation and its role in learning, combined with the advantages of learning analysis, artificial intelligence, and measurement science, providing a choice for researchers in the frontier field of evaluation.

References

[1] J. W. Pellegrino, “The evolution of educational assessment: Considering the past and imagining the future. William H Ang off memorial Lecture,” ETS

[2] P. Griffin and E. Care, “Assessment and teaching of 21st-century skills: Methods and approaches,” Dordrecht: Springer, vol.2, **(2015)**

[3] K. Tremblay, D. Lalancette, and D. Roseveare, “Assessment of higher education learning outcomes (AAHELO): Feasibility study report, design, and implementation,” Paris, France: Organization for Economic Cooperation and Development, **(2012)**

[4] OECD., “The future of education and skills: Education 2030, Geneva, Switzerland,” **(2018)**

[5] S. K. Milligan, G. Kennedy, and D. Israel, “Assessment, credentialing and recognition in the digital era: Recent developments in a fertile field,” Seminar Series272, Centre of Strategic Studies, Melbourne, **(2018)**

[6] S. E. Dreyfus and H. L. Dreyfus, “A five-stage model of the mental activities involved in directed skill acquisition

[7] P. Griffin, “The comfort of competence and the uncertainty of assessment,” Studies in Educational Evaluation, vol.33, no.1, pp.87-99, **(2007)**

[8] S. Messick, "Standards of validity and the validity of standards in performance assessment," Educational Measurement: Issues and Practice, vol.14, no.4, pp.5-8

[9] "Asia-Pacific education research institutes network regional study on transversal competencies in education policy and practice," UNESCO, Bangkok. And Paris. Retrieved from: unesdoc.unesco.org/images/0023/002319/231907E.pdf, **(2015)**

[10] World Economic Forum., The new vision for education: Unlocking the potential of technology. Geneva, Switzerland, **(2015)**

[11] J. D. Bransford, J. D. Brown, and R. R. Cocking, "How people learn: Brain, mind, experience, and school," Expanded edition, Washington DC," National Academy Press. Retrieved from http://www.nap.edu/read/9853/chapter/1, **(2003)**

[12] S. K. Milligan and P. Griffin, "Understanding learning and learning design in MOOCs: A measurement-based interpretation," Journal of Learning Analytics, Special Section on Learning Analytics for 21st Century Competencies: UTS, Australia. Retrieved from https://www.researchgate.net/publication/308272525_Understanding_Learning_and_Learning_Design_in_MOOCs_A_Measurement-Based_Interpretation, **(2016)**

[13] M. Scardamalia, J. Bransford, B. Kozma, and E. Quellmalz, "New assessments and environments for knowledge building," In P. Griffin, B. McGaw, and E. Care (Eds.), Assessment and teaching of 21century skills, New York: Springer, vol.1, pp.231-300, **(2013)**

[14] L. Corrin and P. de Barba, "Exploring students 'interpretation of feedback delivered through learning analytics dashboards. In B. Hegarty, J. McDonald, and S.-K. Loke (Eds.), Rhetoric and Reality: Critical perspectives on educational technology(pp.629-633). Proceedings ASCILITE Dunedin, **(2014)**

[15] R. Luckin, "Towards artificial intelligence-based assessment systems," Nature Human Behaviour,1(0028). Retrieved from https://www.nature.com/articles/s41562-016-0028, **(2017)**

[16] I. Roll and R. Wylie, "Evolution and revolution in artificial intelligence in education," International Journal of Artificial Intelligence in Education, vol.26, no.2, pp.582-599, **(2016)**

[17] J. He, B. I. P. Rubinstein, J. Bailey, R. Zhang, S. Milligan, and J. Chan, "MOOCs meet measurement theory: A topic-modeling approach," Paper presented at the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), Phoenix, Arizona, **(2016)**

[18] M. Flor, S. Y. Yoon, J. Hao, L. Liu, and A. von Davier, "Automated classification of collaborative problem-solving interactions in simulated science tasks," In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp.31-41, **(2016)**

[19] C. Carmean and P. Mizzi, "|The case for nudge analytics," EDUCAUSE Quarterly, Retrieved from https://eric.ed.gov/?id=EJ909992, vol.333, **(2010)**

[20] D. Gasevic, S. Dawson, and G. Siemens, "Let's not forget: Learning analytics is about learning," Tech Trends, vol.59, no.1, pp.64-71, **(2015)**

[21] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," EDUCAUSE Review, vol.46, no.5, pp.30-32, **(2011)**

[22] W. Greller and H. Draschler, "Translating learning into numbers: A framework for learning analytics," Educational Technology and Society, vol.15, no.3, pp.42-47, **(2012)**

[23] W. Greller and H. Draschler, "Translating learning into numbers: A framework for learning analytics," Educational Technology and Society, vol.15, no.3, pp.42-47, **(2012)**

[24] T. J. Cleary, G. I. Callan, and B. J. Zimmerman, "Assessing self-regulation as a cyclical, context-specific phenomenon: Overview and analysis of SLR Microanalytic protocols," Educational Research International. Retrieved from https://doi:10.1155/2012/428639, **(2012)**

[25] R. lterman, "Understanding promotions in a case study of student blogging," Paper presented at the Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium. **(2013)**

[26] J. He, B. I. P. Rubinstein, J. Bailey, R. Zhang, S. Milligan, and J. Chan, "MOOCs meet measurement theory: A topic-modeling approach," Paper presented at the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), Phoenix, Arizona, **(2016)**

[27] J. Hattie and H. Timperley, "The power of feedback," Review of Educational Research, vol.77, no.1, pp.81-112, **(2007)**

[28] S. Buckingham and R. D. Crick, "Multimodal and 21st-century skills learning analytics and datasets," Journal of Learning Analytics, vol.3, no.2, pp.6-21, **(2016)**

[29] V. Shute and M. Ventura, "Stealth assessment: Measuring and supporting learning in video games. Cambridge," MA: MIT Press, **(2013)**

[30] M. Wilson, K. Scalise, and P. Gochyyev, "Assessment of learning in digital interactive social networks: A learning analytics approach," Online Learning, vol.20, no.2, pp.97-119, **(2016)**

[31] S. T. Polyak, A. von Davier, and K. Peterschmidt, "Analyzing game-based collaborative problem solving with computational psychometrics," In Proceedings of ACM KDD conference," Halifax, Nova Scotia, Canada. **(2017)**

[32] M. Wilson, "Constructing measures: An item response modeling approach," New York: Taylor and Francis Group, **(2005)**

[33] R. J. Mislevy and G. D. Haertel, "Implications of evidence-centered design for educational testing," Educational Measurement: Issues and Practice, vol.25, no.4, pp.6-20, **(2006)**