Psychological Test and Assessment Modeling, Volume 65, 2023 (1), 179-190

Machine Learning and Deep Learning in Assessment

Hong Jiao, Qiwei He, Lihua Yao

Background of the Special Issues

Computer-based assessment (CBA) has been dramatically boosted since the onset of the pandemic. The digital assessment environment enables the collection of non-traditional assessment data such as process data, textual data, image data, keystrokes, audio and video data from both traditional assessment platforms and innovative assessment platforms incorporating augmented reality (AR) and virtual reality (VR) technology. Furthermore, process data in addition to item responses in CBA can be easily collected in the digital assessment process. Examples of process data in CBA include item response time, key-stroke, eye-tracking data, action sequence, and answer change behaviors.

Process data may bring new perspectives to better understand the assessment products or accuracy and the process how an item product was attained (Jiao, He, & Veldkamp, 2021). The analyses of these non-conventional structured or unstructured process data call for new methodology other than latent trait modeling to extract more information that the traditional data and analysis methods could not provide. The emergence of big data from a variety of new sources brings ample opportunities and challenges to the traditional assessment framework, which arouses wide attention in interdisciplinary research and real practice.

von Davier, Mislevy and Hao (2021) recently proposed computational psychometrics for analyzing data in digital learning and assessment. They showcased a new methodological perspective using artificial intelligence (AI) methods including supervised and unsupervised machine learning algorithms (Hao & Ho, 2019), deep learning algorithms including Deep Neural Network, Convolutional Neural Network, and Recurrent Neural Network in analyzing multimodal data, time series and stochastic process methods in interactive learning and assessments, social networks analysis, and natural language processing (NLP) for text mining and automated scoring. Over the years, machine learning and deep learning algorithms have been successfully used in automated scoring (e.g., Cummins et al, 2016) and further explored in providing diagnostic feedback to test-takers in writing assessment (e.g., Foltz, 2004; Guo et al., 2018). Recently machine learning algorithms have been explored for cheating detection (e.g., Kim et al., 2016; Liao et al., 2021; Man et al., 2019; Zhou & Jiao, 2022;

Zopluoglu, 2019) and cognitive diagnosis in assessment (e.g., Liu & Cheng, 2018; Jiao et al., 2021). However, the values added from other data sources using the new methodology deserve further extensive exploration. Our special issues hereby called for more applied empirical research and new methods from machine learning and deep learning in analyzing text, image, audio, and video multimodal assessment data. Before highlighting the needs of these special issues, we present a brief overview of machine learning, deep learning, and the use cases of these methodologies that have been demonstrated in previous studies..

Overview of Machine Learning and Deep Learning

Machine learning, as the name suggests, means that machine completes some tasks after learning like human beings. Machine learning is under the bigger umbrella of AI. According to Copeland (2016), deep learning is a subset of machine learning, which is a subset of AI. Since its early work in 1950s, AI developed slowly for over 30 years. In 1980s, machine learning started to grow with the booming use of internet in daily life. Detection of spam emails is a good example of using machine learning algorithms. Starting in 2010s, deep learning made breakthrough in processing image data thanks to the affordable GPUs for speedy parallel processing in analyzing image data, text data, and big data in the digitalized world.

Some well-known use cases emerged in fields like computer vision, speech recognition, text generation, search engine, intelligent assistants, autonomous systems, and robotics. All these advances brought about new impact on human daily life, the world, and re-genesis related to humanities, healthcare, education, and sustainability. Some high-impact success of AI applications includes Alpha Fold (DeepMind), Search and Recommendation Engines (Google), self-driving cars (Tesla) and ChatGPT (Open AI), the latest AI heat.

When the buzz word, AI, is highly utilized in marketing or showcases the advance in the technology of using machine to replace human intelligence, the root of such advances is still machine learning. When the traditional data analysis methods are more related to statistical modeling of numerical structured data with strong assumptions, machine learning algorithms can tackle both structured and unstructured image and text data with pre-processing of image and text data into numerical values in a more naturalistic way. Thus, machine learning is also an interdisciplinary science integrating computer vision processing, computational linguistics, and statistics. On the other hand, deep learning mimics the neurons in the neural networks like human brain. At present, it is not clear how human brain process information within seconds using the neural networks in the brain, it is also hard to understand what is going on when deep learning algorithms process data. Interested readers can refer to different sources for details for the algorithms such as von Davier, Mislevy and Hao (2021) and Hao and Ho (2019).

Use Cases of Machine Learning and Deep Learning in Assessment

One of the most successful applications of machine learning and deep learning in assessment is automated scoring. After the development of the first automated essay scoring system by Page (1966), automated scoring becomes one of the hottest areas that attract researchers from different disciplines including assessment technology, psychometrics, computer science, statistics, computational linguistics, and data science in general. Almost every testing company develops its own in-house proprietary automated scoring engine. These include the "e-rater" developed by Educational Testing Service in 1998 (Attali & Burstein, 2006), the Intelligent Essay Assessor (IEA) developed by Pearson Knowledge and Technologies (Zupanc & Bosnic, 2015), IntelliMetric by Vintage Learning, Bookette by CTB, CRASE by Pacific Metrics, and AutoScore by the American Institute of Research. Over the last two decades, more prominent automated scoring engines were developed including an essay scoring engine developed by Pacific Metrics in its participation in the competition funded by the Hewlett Foundation Automatic Student Assessment Prizes (ASAP). Built upon Page's pioneering work, Measurement Inc. has been developing the PEG system and won the Grand Prizes in the recent Automated Scoring Challenge for the Nation's Report Card (NCES, 2022) for automated scoring of short-answer reading items. More recently, Lottridge (2022) demonstrated the use of transformer neural networks for automated scoring.

Recently, another hot line of research using machine learning and deep learning in assessment is cheating detection/abnormal responding behaviors/item pre-knowledge detection (e.g., Cavalcanti et al., 2012; Gorgun & Bulut, 2022; Hao & Li, 2022; Hao & Fauss, 2022; Kim et al. 2016; Liao et al, 2021; Man et al. 2019; Pan et al., 2022; Pan & Wollack, 2021, 2022; Ranger et al., 2022; Thomas, 2016; Tiong & Lee, 2021; Yan et al., 2022; Zhou & Jiao, 2022 a, b; Zopluoglu, 2019). When both product and process data are available in assessment, cheating or aberrant responding behavior detection which relies on multiple data sources impose challenges on the current psychometric modeling and analysis approaches. Many researchers recently explored using supervised and unsupervised machine learning algorithms and deep learning algorithms for such detection. Further, the recent release of generative AI app, ChatGPT, drew much attention to potential cheating using generative AI in assessment (e.g., Yan et al., 2022).

Further, other researchers studied problem-solving strategy in large-scale assessments by analyzing process data with machine learning or deep learning algorithms (e.g., He et al., 2019; 2021; Tang et al., 2020, 2021). These studies (e.g., Han et al., 2019; Hao et al., 2015; He & von Davier, 2015; 2016; Liao et al., 2019; Stadler et al., 2019) demonstrated the use of machine learning in feature extraction from high-dimensional complex process data. The others (e.g., He et al., 2021, 2022; Jiang et al., 2022; Ulitzsch et al., 2021; Ulitzsch et al., 2022a, 2022b) developed new dynamic sequence mining methods to explore respondents' testing behaviors in interactive tasks.

Some other use cases of machine learning and deep learning in assessment include automated item generation (e.g., Gierl et al., 2012; von Davier, A. et al., 2022; von Davier, M, 2019), estimating item parameters using BERT model or other text features (e.g., Tan et al., 2023; Yancey et al., 2022), enemy item detection (e.g., Chiang & Peabody, 2023b; Fu & Han, 2023; Liu et al., 2023), equating (e.g., Jiang et al., 2022), growth modeling (e.g., Tang & Li, 2022), cognitive diagnosis (e.g., Liu & Cheng, 2018; Jiao et al., 2021), adaptive testing (e.g., Bulut, 2022; Lan, 2022). Further, researchers also explored using machine learning to assess next generation science learning (Zhai, 2022). More recent advances include using deep learning or natural language processing to evaluate construct representation and dimensionality of item pools (e.g., Chiang & Peabody, 2023a), conducting differential item functioning (DIF) analysis (e.g., Mangino et al., 2023) and identifying the causes for DIF (Hoover et al., 2023), shortening an instrument for a targeted screening accuracy (Cheng, 2023), evaluating or collecting validity evidence using NLP (Bulut et al., 2023) or topic modeling methods (Li, 2023), and field testing items using NLP with transformers (Maeda, 2023). We strongly believe this list will keep growing, maybe exponentially in the near future.

What Special Issues 1 and 2 Covered

We co-edited two special issues on machine learning and deep learning in assessment. Among the 11 published papers, two papers (Jung et al., 2022; Ormerod, 2022) studied automated scoring of both essays and short-answer questions. Jung et al (2022) focused on using artificial neural networks for automated scoring of constructed-response items. Ormerod (2022) explored the feature-based interpretability in the developed automated essay scorer using the DeBERTa models. Five papers focused on cheating detection. Gorgun and Bulut (2022) utilized anomaly detection methods to identify aberrant item responses in intelligent tutoring systems. They explored six unsupervised anomaly detection methods including Gaussian Mixture model, Bayesian Gaussian Mixture Model, Isolation Forest, Mahalanobis Distance, Local Outlier factor, and Elliptic Envelope. Pan et al. (2022) proposed a new approach to detect item compromise and preknowledge in computerized adaptive testing built upon the ensemble learning idea. Support Vector Machine (SVM) and a self-training algorithm were used the base models. Using the autoencoder algorithm, a confidence score was adapted for CAT. Zhou and Jiao (2022) explored data augmentation using anomaly detection methods in cheating detection. Tang et al. (2023) explored the LSTM in detecting atypical test-taking behaviors. Yan et al. (2023) investigated detection of GPT-3 generative answers in large-scale high-stakes test. Tang and Li (2022) demonstrated how to use XGBoost models with SHAP credit assignment to calculate student growth percentile, an index often used to track student growth in state accountability system. Zu et al. (2023) presented automated distractor generation for Fill-in-the-Blank vocabulary items using generative AI. The study by He et al (2023) developed two machine learning models: random forest and SVM based multiclass hierarchical classification approaches to predicting problem-solving proficiency levels using process data. Kara et al. (2022) explored prediction of oral reading fluency scores using between-word silence times using NLP and random forest algorithm. Furthermore, model selection for latent Dirichlet allocation in analyzing assessment data was investigated by Mardones-Segovia et al. (2023).

These two special issues are in no means to exhaust the capacity of machine learning and deep learning in providing feasible solutions to issues and challenges in assessment. However, the papers published in these two issues showcased the potentials and promises that machine learning and deep learning algorithms can bring about to improve the current assessment theory and practices in different assessment settings: low-stakes and high-stakes.

Further Exploration

Due to the timeline and the space limits, some important issues related to the applications of machine learning and deep learning in assessment are not addressed, including the interpretability and validity of the results from machine learning and deep learning. In particular, when more advanced deep learning models are used, it would become harder to clearly explain how the input data lead to the output results from the model. Though one paper in special issue 1 (Ormerod, 2022) addressed this issue by exploring mapping between the features and hidden states to facilitate the interpretation of the automated essay scores using the DeBERTa model, it is far beyond enough to inform assessment researchers and practitioners to fully understand the black box. Thus, the validity of using the methods still awaits further exploration.

Further, another important issue in such applications is fairness. No paper in these two special issues addressed this topic. The editors believe that in real applications, model invariance can be checked to assure the fair treatment in developing a population model vs group-specific models. To address this issue, the editors hereby propose differential feature functioning to facilitate the fair interpretation of the results from machine learning or deep learning models, thus enhance the validity and fairness in using the results from machine learning or deep learning models in assessment.

Acknowledgement of the Reviewers

All in all, the general theme of these two special issues focuses on the applications of machine learning and deep learning algorithms in solving psychometric issues and challenges in psychological and educational assessment. We would like to thank all the authors in contributing to the issues. Special thanks go to the reviewers for conducting their reviews and sharing their insights and constructive feedback to authors in a timely fashion. All reviewers are listed below in an alphabetical order:

Bezirhan, Ummugul	Bulut, Okan	Choi, Jaehwa
Choi, Ikkyu	Chung, Jia-Ru	Dworak, Elizabeth
Flanagan, Cathal	Flor, Michael	Gorgun, Guher
Jung, Ji Yoon	Kim, Sungyeun	Lan, Andrew
Liao, Dandan	Lottridge, Susan	Luo, Xin
Luo, Yong	Ormerod, Christopher	Pan, Yiqin
Patten, Jeffrey	Qiao, Xin	Reckase, Mark
Tang, Steven	Tenison, Caitlin S	Ulitzsch, Esther
Wilson, Mark	Woo, Ada	Yan, Duanli
Zhang, Mo	Zhang, Susu	Zhang, Todd (Xing)

Editors
Hong Jiao, University of Maryland, College Park, USA
Qiwei He, Educational Testing Service, USA
Lihua Yao, Northwestern University, USA

References

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment*, 4(3). Available from http://www.jtla.org

Bulut, O. (2022, Nov.). *From Adaptive Testing to Personalized Adaptive Testing: Applications of Machine Learning Algorithms*. Presentation at the 2022 Maryland Assessment Research Center virtual conference.

Bulut, O., MacIntosh, A., & Walsh, C. (2023, April). *Utilizing NLP techniques for collecting validity evidence for a situational judgement test*. Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

Castelvecchi, D. (2016, October 05). *Can we open the black box of AI?* (Nature) Retrieved from https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731

Cavalcanti, E. R., Pires, C. E., Cavalcanti, E. P., & Pires, V. F. (2012). Detection and evaluation of cheating on college exams using supervised classification. *Informatics in Education*, 11(2), 169–190.

Cheng, Y. (2023, April). *How many items we need for 90% screening accuracy with machine learning*? Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

Chiang, Y. C. & Peabody, M. R. (2023a, April). *An NLP approach to evaluating construct representation and dimensionality of item pools*. Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

Chiang, Y. C. & Peabody, M. R. (2023b, April). *Exploring the effects of text-preprocessing methods in enemy item detection*. Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

Copeland, M. (2016). What is the difference between artificial intelligence, machine learning, and deep learning? Retrieved from https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Cummins, R., Zhang, M., & Briscoe, E. (2016, August). *Constrained multi-task learning for automated essay scoring*. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 789–799, Berlin, Germany.

Devlin, J., Chang, M-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding.* arXiv. https://doi.org/10.48550/arXiv.1810.04805

Foltz, P. (2014). *Improving student writing through automated formative assessment: Practices and results*. Proceedings of the 2014 International Association for Educational Assessment Annual Conference (pp. 1-10).

Fröhling, L. & Zubiaga, A. (2021). Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. PeerJ Computer Science. http://dx.doi.org/10.7717/peerj-cs.443

Fu, Y & Han, K. T. (2023, April). *Enemy item detection for quantitative item type*. Presentationat the annual conference of National Council on Measurement in Education. Chicago, IL.

Gao, Y., Cui, Y., Bulut, O., Zhai, X., & Chen, F. (2022). Examining adults' web navigation patterns in multi-layered hypertext environments. *Computers in Human Behavior, 129,* 107142.

Gierl, M., Lai, H., & Turner, S. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*.

Gorgun, G. & Bulut, O. (2022). Identifying aberrant responses in intelligent tutoring systems: an application of anomaly detection methods. *Psychological Test and Assessment Modeling,* 64(4), 359-384.

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement,* 55(2), 194-216.

Han, Z., He, Q., & von Davier, M. (2019). Predictive Feature Generation and Selection Using Process Data from PISA Interactive Problem-Solving Items: An Application of Random Forests. *Frontiers in Psychology, 10,* 1421.

Hao, J. & Fauss, M., (2022, November). *Test security in remote testing age: perspectives from process data analytics and AI*. Presentation at the 2022 annual Maryland Assessment Research Center virtual conference.

Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of Scikit-learn package in Python programming language. *Journal of Educational and Behavioral Statistics, 44*(3), 348–361.

Hao, J. & Li, (2022). *Detecting remote computer access using AI and clickstream Data*. Presentation at the 2022 Conference on Test Security. Princeton, NJ.

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining, 7*(1), 33–50.

He, P. Liu, X., Gao, J., & Chen,W. (2021). *DeBERTa: Decoding-enhanced BERT with distangled attention*. arXiv. https://arxiv.org/pdf/2006.03654.pdf

He, Q., Shi, Q., & Tighe, E. L. (2023). Predicting problem-solving proficiency with multiclass hierarchical classification on process data: A machine learning approach. *Psychological Test and Assessment Modeling*, 65(2), 145-177.

He, Q., Borgonovi, F. & Paccagnella, M. (2019). Using process data to understand adults' problem-solving behaviour in the Programme for the International Assessment of Adult Competencies (PIAAC): Identifying generalised patterns across multiple tasks with sequence mining. *OECD Education Working Papers, No. 205,* OECD Publishing, Paris, https://doi.org/10.1787/650918f2-en.

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Identifying generalized behavioral patterns with sequence mining. *Computers & Education, 166*, 104170. https://doi.org/10.1016/j.compedu.2021.104170.

He, Q., Borgonovi, F., Suárez-Álvarez, J. (2022). Clustering Sequential Navigation Patterns in Multiple-Source Reading Tasks with Dynamic Time Warping Method. *Journal of Computer-Assisted Learning*. 1–18.

He, Q., & von Davier, M. (2015). Identifying Feature Sequences from Process Data in Problem-Solving Items with N-grams. In A. van der Ark, D. Bolt, S. Chow, J. Douglas & W. Wang (Eds.), *Quantitative Psychology Research: Proceedings of the 79th Annual Meeting of the Psychometric Society* (pp.173–190). New York: Springer.

He, Q., & von Davier, M. (2016). Analyzing Process Data from Problem-Solving Items with N-Grams: Insights from a Computer-Based Large-Scale Assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.) *Handbook of Research on Technology Tools for Real-World Skill Development* (pp. 749–776). Hershey, PA: Information Science Reference.

Hoover, J., Thompson, W. J., & Frey, B. (2023, April). *Using machine learning to identify causes of differential item functioning*. Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

Jiao, H., He, Q., Veldkamp, B. (2021, eds.). *Process data in educational and psychological measurement*, Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-138-0

Jiao, H., Zhou, T, & Ding, Y. (2021, July). *Analyzing responses, response time and answer changes for cognitive diagnosis using machine learning algorithms*. Presentation at the 2021 Virtual International Meeting of the Psychometric Society.

Jiang, Y., Cayton-Hodges, G. A., Oláh, L. N., & Minchuk, I. (2023). Using sequence mining to study students' calculator use, problem solving, and mathematics achievement in the National Assessment of Educational Progress (NAEP). *Computers & Education, 193,* 104680.

Jiang, Z., Han, Y., Xu, L., Shi, D., Liu, R., Ouyang, J., & Cai, F. (2022). The NEAT equating via chaining random forests in the context of small sample sizes: A machine-learning method. *Educational and Psychological Measurement*, https://doi.org/10.1177/00131644221120899

Jung, J., Tyack, L., & von Davier, M. (2022). Automated scoring of constructed-response items using artificial neural networks in international large-scale assessment. *Psychological Testing and Assessment Modeling.* 64(4), 471-494.

Kara, Y., Kamata, A., Ozkeskin, E. E., Qiao, X., & Nese, J. F. T. (2023). Predicting oral reading fluency scores by between-word silence times using natural language processing and random forest algorithm. *Psychological Test and Assessment Modeling*, 65(2), 36-54.

Kim, D., Woo, A., & Dickison, P. (2016). Identifying and investigating aberrant responses using psychometrics-based and machine learning-based approaches. In *Handbook of Quantitative methods for detecting cheating on tests* (pp. 70–97). Routledge.

Kohonen, T. (1997). Self-Organizing Maps. Heidelberg: Springer-Verlag. doi: 10.1007/978-3-642-97966-8.

Lan, A. (2022, November). *Data-driven item selection and generation in assessments and learning*. Presentation at the 2022 Maryland Assessment Research Center virtual conference.

Li, J. (2023, April). *Evaluating content-related validity of mathematical diagnostic items using a topic modeling approach*. Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

Liao, D., He, Q., & Jiao, H. (2019). Mapping background variables with sequential patterns in problem-solving environments: An investigation of U.S. adults' employment status in PIAAC. *Frontiers in Psychology, 10*, 646.

Liao, M., Patton, J., Yan, R., & Jiao, H. (2021). Mining process data to detect aberrant test takers. *Measurement: Interdisciplinary Research and Perspectives*. 19(2), 93-105. https://doi.org/10.1080/15366367.2020.1827203

Liu, C., & Cheng, Y. (2018). An Application of the Support Vector Machine for Attribute-By-Attribute Classification in Cognitive Diagnosis. *Applied Psychological Measurement*. 42(1), 58-72. doi: 10.1177/0146621617712246.

Liu, L., Walker, M., & Weir, J. (2023, April). *An exploration of natural language processing for enemy item detection*. Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

Lottridge, S. (2022, November). *Applications of transformer neural networks in processing examinee responses*. Presentation at the 2022 Maryland Assessment Research Center virtual conference.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (pp. 281–297). University of California Press. https://projecteuclid.org/euclid.bsmsp/1200512992

Maeda, H. (2023, April). *Field-Testing items using artificial intelligence: Natural Language Processing with transformers*. Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

Man, K., Harring, J. R., & Sinharay, S. (2019). Use of data mining methods to detect test fraud. *Journal of Educational Measurement*, 56(2), 251-279.

Mangino, T. A., Finch, H., French, B., & Demir, C. (2023, April). *Identification of differential item functioning using machine learning*. Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

Mardones-Segovia, C., Wheeler, J. M., Choi, H., Wang, S., & Cohen, A. S. (2023). Model selection for latent Dirichlet allocation in assessment data. *Psychological Test and Assessment Modeling*, 65(2), 3-35.

Olah, C. (2015, August 27). *Understanding LSTM Networks*. Retrieved from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Ormerod, C. M. (2022). Mapping between hidden states and features to validate automated essay scoring using DeBERTa models. *Psychological Testing and Assessment Modeling*. 64(4), 495-526.

Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Page, E.B. (1968). The use of the computer in analyzing student essays, *International Review of Education*, 14(3), 253-263

Pan, Y., Sinharay, S., Livne, O., & Wollack, J. A. (2022). A machine learning approach for detecting item compromise and preknowledge in computerized adaptive testing. *Psychological Test and Assessment Modeling*. 64(4), 385-424.

Pan, Y., & Wollack, J. A. (2021). An unsupervised-learning-based approach to compromised items detection. *Journal of Educational Measurement*, *58*(3), 413-433.

Pan, Y., & Wollack, J. A. (2022). *An ensemble-unsupervised-learning-based approach for the simultaneous detection of preknowledge in examinees and items when both are unknown*. https://doi.org/10.31234/osf.io/jtr78.

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology*. 9:2231. doi: 10.3389/fpsyg.2018.02231.

Ranger, J., Schmidt, N., Wolgast, A. (2022). Detecting cheating in large-scale assessment: the transfer of detectors to new tests. *Educational and Psychological Measurement*. Online First. https://doi.org/10.1177/00131644221132723

Stadler, M., Fischer, F., & Greiff, S. (2019). Taking a closer look: An exploratory analysis of successful and unsuccessful strategy use in complex problems. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.00777.

Tan, R. B., Bulut, O., Gorgun, G., & Wongvorachan, T. (2023, April). *Mining textual features of questions to predict item parameters*. Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

Tang, S., & Li. Z. (2022). Considerations in using XGBoost models with SHAP credit assignment to calculate student growth percentiles. *Psychological Test and Assessment Modeling*, 64(4), 445-470.

Tang, S., Samuel, S., & Li. Z. (2023). Detecting atypical test-taking behaviors with behavior prediction using LSTM. *Psychological Test and Assessment Modeling*, 65(2), 76-124.

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika, 85*, 378-397.

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology*, *74*(1), 1-33.

Thomas, S. L. (2016). *The use of data mining techniques to detect cheating*. Presentation at the 2016 Maryland Assessment Research Center conference. College Park, MD.

Tiong, L., & Lee, H. (2021). E-*cheating prevention measures: Detection of cheating at online examinations using deep learning approach -- A case study*. arXiv:2101.09841v1.

Ulitzsch, E., He, Q., & Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics, 47*(1), 3-35.

Ulitzsch, E., He, Q., Ulitzsch, V., Nichterlein, A., Molter, H., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika*, 1–25.

Ulitzsch, E., Ulitzsch, V., He, Q., & Lüdtke, O. (2022). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods*, 1–21.

von Davier, A. A., Runge, A., Attali, Y., Park, Y., Yancey, K., LaFlair, G., Horie, A. (2022,

Nov.). *The item factory: intelligent automation in support of test development at scale*. Presentation at the 2022 Maryland Assessment Research Center virtual conference.

von Davier, A. A., Mislevy, R. J., & Hao, J. (2021, eds.), *Computational psychometrics: new methodologies for a new generation of digital learning and assessment*, Methodology of Educational Measurement and Assessment, Springer. https://doi.org/10.1007/978-3-030-74394-9_1

von Davier, M. (2019). *Training optimus prime, M.D.: Generating medical certification items by fine-tuning ApenAI's GPT2 transformer model*. https://arxiv.org/abs/1908.08594

Williamson, D. M., Mislevy, R. J., and Bejar, I. I. (2006, eds). *Automated Scoring of Complex Tasks in Computer-Based Testing*, Mahwah, NJ: Lawrence Erlbaum Associates, Publishers. doi: 10.4324/9780415963572

Yan, D., Fauss, M., Hao, J., & Cui, W. (2023). Detection of AI-generated essays in writing assessment. *Psychological Testing and Assessment Modeling.* 65(2), 125-144.

Yan, D., Fauss, M., Cui, & Hao, J. (2022). *Detection of AI Generated Essays*. Presentation at the 2022 Conference on Test Security. Princeton, NJ.

Yancey, K. P., Runge, A., Lockwood, J. R., & LaFlair, G. T. (2022, November). *Estimating contextual word-level item parameters using BERT in an IRT framework.* Presentation at the 2022 annual Maryland Assessment Research Center virtual conference.

Zhai, X. (2022, November). *Using machine learning to assess next generation science learning*. Presentation at the 2022 annual Maryland Assessment Research Center virtual conference.

Zhou, T. & Jiao, H. (2022a). Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644221117193

Zhou, T. & Jiao, H. (2022b). Data augmentation in machine learning for cheating detection: An illustration with the blending learning algorithm. *Psychological Testing and Assessment Modeling*. 64(4), 425-444.

Zopluoglu, C. (2019). Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (XGBoost). *Educational and Psychological Measurement*, 79. doi: 10.1177/0013164419839439.

Zu, J., Choi, I., & Hao, J. (2023). Automated distractor generation for fill-in-the-blank items using a prompt-based learning approach. *Psychological Testing and Assessment Modeling*. 65(2), 55-75.

Zupanc, K., & Bosnic, Z. (2016). Advances in the field of automated essay evaluation. *Informatica*, *39*(4), 383–395.