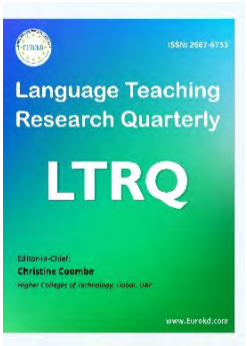

Language Teaching Research Quarterly

2023, Vol. 37, 204–212

Leveraging Computational Psychometrics for Language Testing

Ardeshir Geranpayeh

Founder & Director, Cambridge Computational Psychometrics, United Kingdom

Received *16 May 2023* ***Accepted*** *18 October 2023*

Abstract

The recent surge in the popularity of Large Language Models (LLM) for language assessment underscores the growing significance of cost-effective language evaluation in our increasingly digitalized society. This paper posits that the application of computational psychometrics can enable the incorporation of technology into language assessment, enhancing test accessibility for learners while simultaneously elevating the precision of language proficiency evaluation. In this context, computational psychometrics is defined as a fusion of theory-based psychometrics and data-driven methodologies drawn from machine learning, artificial intelligence, natural language processing, and data science. This amalgamation offers a more robust and adaptable framework for analyzing intricate data, particularly within the contemporary landscape of learner-centric assessment. The paper concludes by emphasizing that the integration of computational psychometrics into language assessment opens up promising avenues for future research and practical applications, heralding an era of innovation in this field.

Keywords: *Computational Psychometrics, Language Testing, Large Language Models*

Introduction

Globalization has highlighted greater need for a much affordable language assessment where technology facilitates the inclusion of new methods which reflect the multimodal constructs of our language assessment in the modern integrated and often multilingual communities. Against this background, I argue that 'computational psychometrics' (CP) can be at the forefront of an interdisciplinary new field which brings about efficiency in test construction processes and provides a much more affordable assessment for learners while addressing new challenges of complex multi-construct nature of language assessment that could not be handled by traditional psychometrics. Computational psychometrics marries the principles of psychometric assessment with machine learning methodologies for assessment and learning purposes. The common ground between these two domains lies in their approach to constructing assessments.

* Corresponding author.
E-mail address: ardeshirgeranpayeh@gmail.com https://doi.org/10.32038/ltrq.2023.37.11

In psychometrics, the process begins with a theory-driven definition of the assessment construct and subsequently gathers evidence to support that theory, often employing methods such as Evidence-Centered Design (ECD) (Mislevy et al., 2003; Mislevy, 2018). This approach is predominantly top-down in nature. Conversely, machine learning takes a different route by commencing with observed assessment data (i.e., candidates' performance) and endeavors to deduce the underlying factors influencing the measured construct(s). This represents a bottom-up methodology. The incorporation of computational psychometrics into assessment yields comprehensive data that captures the multifaceted nature of the new assessment, encompassing interdependencies across intricate tasks and item types.

In recent years computational psychometrics has emerged as a powerful tool in the field of language assessment, revolutionizing the way we measure and evaluate language proficiency. This article delves deeply into the utilization of Computational Psychometrics (CP) within the realm of language assessment, shedding light on the strides achieved in this domain and deliberating on their ramifications for the field of language testing. By leveraging psychometric techniques such as the use of item response theory, computerized adaptive testing, and natural language processing, CP offers improved precision, efficiency, and fairness in language assessment. This article also addresses potential challenges and ethical considerations associated with the use of computational psychometrics. Overall, the integration of CP in language assessment has the potential to enhance the accuracy and validity of language proficiency measurement, providing valuable insights for learners, educators, and policymakers.

Definition

The term computational psychometrics had been used by Tatsuoka and Tatsuoka (2013) in the context of computer multi-stage testing. von Davier (2015), however, was the first person that introduced the term computational psychometrics in a modern sense during an invited presentation at the International Conference of Machine Learning in Lille, France, held in July 2015, and later in a more formal definition, CP was characterized as an amalgamation of data science techniques, computer science, and rigorous psychometric methods. Its purpose is to facilitate the examination of intricate data stemming from performance assessments and technology-enhanced learning and assessment systems (LAS) (von Davier, 2017).

Computational Psychometrics represents an interdisciplinary convergence, uniting the fundamental principles of psychometric assessment with the methodologies of Machine Learning for both assessment and learning purposes. The common thread between these two fields lies in their distinct approaches to constructing assessments. In the realm of Psychometrics, the process commences with a theory-driven definition of the assessment construct, progressing towards the provision of evidence, often through methods like Evidence-Centered Design (ECD) (Mislevy et al., 2003; Mislevy, 2018). This approach is primarily characterized as top-down. In contrast, Machine Learning embarks on a different journey, commencing with the examination of observed assessment data (i.e., candidate's performance) and striving to deduce the underlying factors influencing the measured construct. This represents a bottom-up approach. The intersection of these two disciplines occurs at the juncture of data (observation), which falls somewhere in the middle of their methodologies.

Both fields heavily rely on data science and employ data mining techniques to analyze the data at hand.

Effectively managing intricate data necessitates adopting fresh perspectives and methodologies capable of accommodating the intricacies inherent in the constructs being measured or observed. This adaptation also involves embracing digital technology, which has seamlessly integrated into our daily communication routines, becoming an indispensable component of our lives.

CP harnesses the potential of cutting-edge technology, leveraging it not only for the analysis of performance data but also for the entire spectrum of assessment processes, including item design, administration, test security, and fairness enforcement through digital data utilization. This approach broadens the scope, speed, and volume of accessible data, enabling the introduction of innovative methods and strategies to tackle emerging challenges. These challenges encompass various facets, such as identifying evidence, modeling data, and assessing and predicting learners' performance within intricate scenarios, including collaborative tasks (mediation), interactive tasks, game/simulation-based activities, and multimodal learning and assessment tasks (von Davier et al., 2021).

In brief, computational psychometrics can be succinctly characterized as an amalgamation of theory-based psychometrics and data-driven methodologies originating from machine learning, artificial intelligence, natural language processing, and data science. This fusion collectively offers an enhanced theoretical and practical framework for the examination of intricate data, especially within the contemporary context of learner-centered assessments prevalent in our present era.

Advancements in Computational Psychometrics for Language Assessment

The use of AI and machine learning in language testing is not new but as Xi (2010) reported it had mainly been limited to automated scoring of writing and speaking skills. There has been a special issue of the journal of *Language Testing* (Xi, 2010) dedicated to the topic of Automated scoring and feedback systems for language assessment and learning. There have been multiple publications in the use of various aspects of CP in assessment since (see Cipresso et al., 2015; Flor et al., 2016; Greiff et al., 2017; Marsman et al.,2018; Mislevy, 2018 ; Polyak et al., 2017; von Davier et al., 2017, 2018; to name but a few). There has also been a rapid increase in the use of CP in language assessment since the invention of Generative Pre-training Transformers (GPT). The literature in the latter field is dominated by researchers in Duolingo as they widely use GPT in their language assessment model (see Attali et al., 2022; Burstein et al., 2022; Cardwell et al., 2022; Cardwell et al., 2023; Godwin & Naismith, 2023; Goodwin et al., 2023; Ishikawa & Settles, 2016); Kunnan et al., 2022; LaFLair, 2020; Langenfeld et al., 2022; McCarthy et al., 2021). With the wide application of GPT, the CP is now being used not only in scoring essays but also in Automatic Item Generation (AIG), creation of new item types (Park et al., 2022; Laflair et al., 2023) and even in providing a better test security framework in a digital-first assessment context (LaFlair et al., 2022).

Benefits of CP in Language Assessment

There are many benefits in using CP for language assessment. One obvious area is the improved precision and reliability. By employing new advanced statistical models, CP

enhances the precision and reliability of language proficiency measurement. Item banks developed using IRT facilitate the creation of assessments that precisely target an individual's ability level. This should result in more accurate and trustworthy assessments.

Another advantage is the increased Efficiency and Cost-Effectiveness: Computerized adaptive testing used in CP optimizes the test-taking experience by adapting the test to each individual's proficiency level. As a result, unnecessary or redundant questions are eliminated, reducing test duration. This improved efficiency saves time and resources, making language assessment more cost-effective.

Another benefit of Computational Psychometrics (CP) is the improved promotion of fairness and accessibility in our assessments. CP has the potential to address issues of fairness and accessibility in language assessment in new ways. Adaptive testing ensures that test-takers are presented with items that accurately reflect their ability, preventing both underestimation and overestimation of proficiency. Moreover, computer-based assessments can be delivered remotely, making them accessible to a wider range of individuals, including those with physical disabilities or geographic constraints. This has already proved very effective during the pandemic when thousands of physical test centres had to be shut down. For a review of the fairness of the use of remote proctoring during pandemic see Isbell et al (in press).

Test security enhancement

One of the main advantages of the application of CP in language assessment is its contribution to improving test security in several ways:

Remote Proctoring: CP enables the integration of remote proctoring technologies into online assessments. Remote proctoring systems utilize various methods such as video monitoring, audio monitoring, keyboard strike patterns and screen recording to monitor test-takers during the assessment. These systems help detect and deter cheating behaviors by providing a virtual proctoring presence and capturing any suspicious activities in real-time.

Automated Plagiarism Detection: CP can incorporate automated plagiarism detection algorithms into assessments that involve written responses, such as essays or short answers. These algorithms compare the test-taker's responses to a large database of existing texts to identify any instances of plagiarism. This helps ensure the originality and integrity of the test-taker's work.

Automated Item Generation: Automated item generation (AIG) represents a paradigm in the creation of assessment items or test questions, making use of principles from artificial intelligence and automation. True to its name, AIG aims to streamline and automate a significant portion, if not all, of the laborious item authoring process, a well-known time-consuming aspect of assessment development—familiar to anyone who has authored test questions. Developing individual items can incur costs as high as $2500, underscoring the substantial financial and time investments required. Consequently, even a modest reduction in the average cost could result in significant time and cost savings for organizations

Randomized Item Selection: CP can employ algorithms to randomly select items from an item bank during the test administration process. This randomization reduces the predictability of the test content, making it more difficult for test-takers to share specific item information with others who may take the test later.

Item Exposure Control: CP can implement item exposure control strategies to limit the exposure of test items to test takers. These strategies aim to prevent the widespread dissemination of test items, which could give an unfair advantage to some test-takers. By controlling the exposure of items, CP helps maintain the security and validity of assessments

Data Encryption and Secure Storage: CP emphasizes the importance of data security by employing encryption techniques to protect sensitive test-taker information during transmission and storage. Secure data storage systems with robust access controls are implemented to prevent unauthorized access to test data

Data Forensics: CP can employ data forensics techniques to identify and investigate any irregular patterns or anomalies in test responses that may indicate cheating or fraudulent behavior. Statistical methods and algorithms can be automated to flag suspicious response patterns, ensuring the integrity of the assessment results.

It is imperative to emphasize that while CP provides valuable tools to augment test security, its utilization should be complemented by other security measures and practices to establish a comprehensive and resilient test security framework. To effectively safeguard the integrity of language assessments, it is essential to engage in regular monitoring, ongoing research, and a continuous enhancement of security measures, as outlined by Geranpayeh (in press). This proactive approach is crucial for staying ahead of evolving threats and ensuring the continued integrity of language assessments.

By leveraging computational approaches, CP aims to enhance the precision, reliability, efficiency, security and fairness of language assessment. CP can achieve that by applying computational methods to analyze language-related data, such as responses to language tasks, written essays, or spoken language samples. These methods enable the automatic scoring, analysis of linguistic features, and the development of adaptive assessments that dynamically adjust the difficulty of test items based on the individual's proficiency level. CP can also enhance our understanding and evaluation of psychological constructs, including but not limited to language proficiency, cognitive abilities, personality traits, and mental health indicators.

Overall, computational psychometrics brings together the power of computational techniques (data science) and psychometric principles to advance the field of language assessment, ultimately improving the measurement and understanding of language proficiency.

Applications of CP in Language Assessment

CP can basically be used in three different areas: language proficiency testing, language learning and instruction and language program evaluation.

As we have already argued, CP enables the development of more accurate and comprehensive language proficiency tests. Adaptive assessments tailored to individual abilities provide more precise measurements, allowing for finer-grained distinctions between proficiency levels as well as providing a fairer platform for all test takers.

CP supports personalized language learning by providing detailed diagnostic information about a learner's strengths and weaknesses. This information can guide instructional interventions and facilitate targeted feedback, enhancing the effectiveness of language instruction.

Significant advancements have emerged in the realm of language learning and feedback in recent years. For instance, Yannakoudakis et al. (2018) introduced the development of an automated placement system designed for ESL learners known as Write & Improve (https://writeandimprove.com/). This cloud-based tool is freely accessible and offers automated diagnostic feedback to non-native English language learners, catering to varying levels of detail.

Furthermore, Geranpayeh and Yannakoudakis (2017) presented findings on the diagnostic feedback generated by such an automated learning system, while Geranpayeh and Saville (2017) showcased the advancements and diagnostic feedback observed in real classroom settings.

Another notable application of Computational Psychometrics (CP) is in the evaluation of language programs. CP can evaluate the effectiveness of language programs by scrutinizing large-scale datasets collected from assessments. This evaluation yields valuable insights into program outcomes, pinpointing areas that require improvement and contributing to decision-making based on empirical evidence. Griggs and Crain-Dorough (2021) explored the potential of appreciative inquiry in program evaluation and research.

CP Challenges and Ethical Considerations

While CP offers numerous advantages, there are challenges and ethical considerations that must be addressed.

Test Security and Cheating Prevention: Computer-based assessments raise concerns about test security and cheating prevention. Implementing robust security measures, such as remote proctoring and plagiarism detection algorithms, is essential to ensure the integrity of language assessments. Geranpayeh (in press) has listed a number of ways that CP can help improve test security. LaFlair et al (2022) describe how CP can provide a secure environment for digital-first assessment via human in the loop process.

Privacy and Data Protection: The collection and analysis of large-scale language assessment data require strict adherence to privacy and data protection regulations. Test developers must adopt appropriate measures to safeguard personal information and ensure compliance with relevant legal and ethical standards. The assessment industry is working to set up standards highlighting how privacy and data protection needs to be addressed in a digital environment. There are already guidelines on responsible AI standards (Burstein, 2023).

Bias and Fairness: CP must strive to address bias and fairness issues. Algorithms used in language assessment should be regularly monitored and refined to minimize bias related to gender, race, culture, or socioeconomic status, ensuring fair and equitable assessments for all individuals. Belzak (2023) reports on measuring bias using regularized Differential Item Functioning in a digital-first assessment.

Final Remarks

Computational psychometrics represents a significant leap forward in the realm of language assessment, offering the potential for more precise, efficient, and equitable evaluations of language proficiency. Within the domain of language assessment, computational psychometrics employs advanced statistical models, machine learning algorithms, and natural

language processing techniques to measure and assess language proficiency. This transformative approach enhances various aspects of language assessment, including item calibration, test construction, scoring, and adaptive testing, all facilitated by computational tools and algorithms.

By harnessing methodologies such as item response theory, computerized adaptive testing, and natural language processing, computational psychometrics has the capacity to revolutionize language assessment practices, delivering benefits to learners, educators, and policymakers alike. However, the responsible and equitable implementation of computational psychometrics in language assessment demands careful consideration of challenges and ethical considerations.

Computational psychometrics enables the creation of items that are tailored to measure specific skills or knowledge domains, optimizing the assessment process. Additionally, computational psychometrics assists in the calibration and validation of generated items, ensuring their reliability and fairness in evaluating individual performance. In summary, computational psychometrics has become an invaluable tool in automating item generation (AIG), contributing to the development of more effective and accurate assessments across various domains.

In essence, computational psychometrics amalgamates the capabilities of computational techniques, such as data science, with the foundational principles of psychometrics to propel the field of language assessment forward, ultimately enhancing our ability to measure and comprehend language proficiency.

Moreover, the integration of computational psychometrics into language assessment introduces promising avenues for future research and practical applications. Ongoing advancements in machine learning, artificial intelligence, and natural language processing will continually refine and expand the capabilities of language assessment systems.

ORCID

https://orcid.org/0009-0000-0535-0569

Acknowledgements

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

References

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Natural Language Processing*, 5, https://doi.org/10.3389/frai.2022.903077

Belzak, W.C.M. (2023). The regDIF R package: Evaluating complex sources of measurement bias using regularized differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), 974-984. https://doi.org/10.1080/10705511.2023.2170235

Burstein, J. (2023). Responsible AI standards. *Duolingo White Paper*. Retrieved from https://go.duolingo.com/ResponsibleAI

Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2022). A theoretical assessment ecosystem for a digital-first assessment: The Duolingo English test. *Duolingo Research Report DDR-22-01*.

Cardwell, P. R., LaFlair, G. T., Naismith, B., & Settles, B. (2022). Duolingo English test: Technical manual. *Duolingo Research Report DDR-22-08.*

Cardwell, P. R., Goodwin, S., Naismith, B., LaFlair, G. T., Lo, K. L., & Yancey, K. P. (2023). Assessing Speaking on the Duolingo English Test. *Duolingo Research Report DRR-23-03.*

Cipresso, P., Matic, A., Giakoumis, D., & Ostrovsky, Y. (2015). Advances in computational psychometrics. *Computational and Mathematical Methods in Medicine*, 1–2. https://doi.org/10.1155/2015/418683

Flor M., Yoon S.Y., Hao J., Liu L., & von Davier A.A. (2016). Automated classification of collaborative problem solving interactions in simulated science tasks. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego, CA: Association for Computational Linguistics. http://doi.org/10.18653/v1/W16-0504

Geranpayeh A. (in press). Detecting Plagiarism and Cheating in A.J. Kunnan (Ed.) *The companion to language assessment.* Wiley-Blackwell Publishing.

Geranpayeh A. & Saville N. (2017). Learning and assessment: progress and diagnostic testing in the classroom. Paper presented at *NCME Special conference on classroom assessment and large-scale psychometrics*: the Twin shall meet, Lawrence, KS, 12-14 September 2017.

Geranpayeh A. & Yannakoudakis H. (2017). 'Making the connections: Digital innovation and diagnostic feedback.' Plenary paper presented at the *6th ALTE International conference*, Bologna, Italy, 3-5 May 2017.

Goodwin, S., Attali, Y., LaFlair, G.T., Park, Y., Runge, A., von Davier, A.A., & Yancey, K.P. (2022). Duolingo English Test-Writing. *Duolingo Research Report DRR-22-03*.

Goodwin, S., & Naismith, B. (2023). Assessing listening on the Duolingo English Test. *Duolingo Research Report DRR-23-02.*

Greiff, S., Gasevic, D., & von Davier, A.A. (2017). Using process data for assessment in intelligent tutoring systems: A psychometrician's, cognitive psychologist's, and computer scientist's perspective. *Army Research Laboratory, 171–179*.

Griggs, D., & Crain-Dorough, M. (2021). Appreciative inquiry's potential in program evaluation and research. *Qualitative Research Journal, 21(4), 375-393.*

Isbell, D.R., Kremmel, B., & Kim, J. (in press). Remote proctoring in language testing: Implications for fairness and justice. In Xi (Ed.), *Language Assessment Quarterly*.

Ishikawa, L., Hall, K., & Settles, B. (2016). The Duolingo English test and academic English. *Duolingo Research Report DDR-16-01*.

Kunnan, A.J., Qin, C.Y., & Zhao, C.G. (2022). Developing a Scenario-Based English Language Assessment in an Asian University. *Language Assessment Quarterly*, *19*(4), 368-393. https://doi.org/10.1080/15434303.2022.2073886

LaFlair, G.T. (2020). Duolingo English Test: Subscores. *Duolingo Research Report DRR-20-03*.

LaFlair, T., Langenfeld, T., Baig, B., Horie, A.K., Attali, Y., & von Davier, A.A. (2022). Digital-first assessments: A security framework. *Journal of Computer Assisted Learning*, 1–10. https://doi.org/10.1111/jcal.12665

LaFlair, G.T., Runge, A., Attali, Y., Park, Y., Church, J., & Goodwin, S. (2023). Interactive Listening—The Duolingo English Test. *Duolingo Research Report DRR-23-01*.

Langenfeld, T., Burstein, J., & von Davier, A.A. (2022). Digital-First Learning and Assessment Systems for the 21st Century. *Frontiers in Education*, 7, 857604. https://doi.org/10.3389/feduc.2022.857604

Marsman M., Borsboom D., Kruis J., Epskamp S., van Bork R., Waldorp L.J., van der Maas H.L.J., & Maris G. (2018). An Introduction to network psychometrics: Relating ising network models to item response theory models. *Multivariate Behavioral Research*, 53 (1), 15–35. https://doi.org/10.1080/00273171.2017.1379379

McCarthy A.D., Yancey K.P., LaFlair G.T., Egbert J., Liao M., & Settles B. (2021). Jump-Starting item parameters for adaptive language tests. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 883–899. November 7–11, 2021. c 2021 Association for Computational Linguistics.

Mislevy R.J. (2018). *Sociocognitive foundations of educational assessment.* Routledge.

Mislevy R.J., Steinberg L.S. & Almond R.A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspective,* 1, 3-67.

Park Y., LaFlair G.T., Attali Y., Runge A., & Goodwin S. (2022). Interactive reading – The Duolingo English Test. *Duolingo Research Report DRR-22-02.*

Polyak S.T., von Davier A.A., & Peterschmidt K. (2017). Computational Psychometrics for the Measurement of Collaborative Problem Solving Skills. *Frontiers in Psychology*, 8, 20–29. https://doi.org/10.3389/fpsyg.2017.02029

Tatsuoka C. & Tatsuoka M.K. (2013). Computational psychometrics in large-scale assessment programs. In *Computerized multistage testing* (pp. 49-68*)*. Chapman and Hall/CRC.

von Davier A.A. (2015). Computational psychometrics. Invited presentation at the *pre-conference workshop "Machine Learning in Education"* at the International Conference of Machine Learning, Lille, France.

von Davier A.A. *(*2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, 54 (1), 3–11. https://doi.org/10.1111/jedm.12129

von Davier A.A., Mislevy R.J., & Hao J. (Eds) (2021). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python.* Springer Nature.

von Davier A.A., Hao J., & Kyllonen P.C., (2017). Interdisciplinary research agenda in support of assessment of collaborative problem solving: lessons learned from developing a Collaborative Science Assessment Prototype. *Computers in Human Behavior*, 76, *631–640.* https://doi.org/10.1016/j.chb.2017.04.059

von Davier A.A., Liu L., Hao J., Yoon, S.Y., & Flor M. (2018). "Automated classification of collaborative problem solving interactions in simulated science tasks ". *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications:* 31–41. https://doi.org/10.18653/v1/W16-0504*. S2CID 390510 – via aclanthology.coli.uni-saarland.de.*

von Davier A.A., Zhu M., & Kyllonen P.C. (2017). *Innovative assessment of collaboration (1 ed.).* Springer International Publishing.

Xi, X. (2010). Automated scoring and feedback systems: Where are we heading? *Language Testing*, 27 (3), 291-300. https://doi.org/10.1177/0265532210364643

Xi X. (Ed.). (in press). *Language Assessment Quarterly* special issue on the role of AI and language assessment.

Yannakoudakis H., Andersen Ø., Geranpayeh A., Briscoe T., & Nicholls D. (2018). Developing an automated writing placement system for ESL learners. *Applied Measurement in Education, 31* (3), 251-267. https://doi.org/10.1080/08957347.2018.1464447