A Consideration of Factors Affecting the Use of Automatic Item Generation (AIG) in Developing Items for Use in Certification and Licensure Assessments: *A Review of the Literature*

GEMMA CHERRY, Ph.D., Prometric Post-Doctoral Researcher in Assessment
CONOR SCULLY, MSc, Prometric Ph.D. Candidate
MICHAEL O'LEARY, Ph.D., Director
Centre for Assessment, Research, Policy and Practice in Education (CARPE)
Dublin Center University
&
LINDA WATERS, Ph.D., Vice President, Prometric

Introduction

Test questions/items are essential for assessing learners' knowledge and/or understanding (Ch & Saha, 2020). However, item construction using traditional, manual methods can be challenging and is a process that involves many trained experts including item writers, subject experts, and psychometricians (Kurdi et al., 2020). It is also fundamental that the items developed are of a high quality and that there is a large enough pool from which to draw (Kurdi et al., 2020). This is because items frequently need to be replaced to ensure test security and minimise item exposure (Gierl & Lai, 2012; Kurdi et al., 2020). As such, manual test item generation can prove to be an expensive and time-consuming task (Ch & Saha, 2020).

Automatic Item Generation (AIG) has emerged in recent decades as a response to the challenges involved with traditional item development. AIG is a rapidly evolving research area where various methods are used to generate items using computer technology and item models[1] based on a set of rules (algorithms) (Gierl & Lai, 2012; Gierl, Ball, et al., 2015). AIG is thought to provide a method for the continuous production of high volumes of content-specific test items (Gierl & Lai, 2013). Von Davier (2018) highlights that because the majority of items used in commercial testing continue to be written by humans, AIG may prove to be a valuable resource in terms of enhancing and shortening the overall item development process. However, to date, much of the research and literature focused on AIG has taken place in the context of K-12 and university level education. The purpose of this paper is to provide an overview of current trends in AIG and detail its applicability and relevance for large-scale assessments in certification and licensure. This paper presents the key findings from a literature review focusing on the critiques and benefits of using AIG. A series of recommendations is then provided to aid decision making regarding incorporating AIG into the item development process. The paper concludes by highlighting several areas that are deserving of future research.

Approaches to Automatic Item Generation

There are numerous different approaches to AIG that have been identified in the literature. Kurdi et al. (2020) conducted a systematic review of over 70 papers and determined that approaches to AIG could be classified along two different dimensions: *level of understanding* and *procedure of transformation.*

- The *level of understanding* dimension determines the extent to which test developers need to have comprehension of the meaning of the text/information that is being used to automatically generate the

[1] An item model, as described by Gierl, Lai, and Turner (2012), is a template that contains the information (e.g., item stem variations, elements of the stem that can be manipulated, response options) needed to generate test items. This information can be manipulated, thus enabling a single item model to generate a large number of unique items. For further detailed information, please see Gierl et al. (2012).

items. Within this dimension are **syntactic** and **semantic-based** approaches to AIG. In the syntactic approach, items are generated automatically from a given piece of text. Under the syntactic approach, test developers are not required to have a working understanding of the meaning of the text itself. In light of this, the role of humans in the item development process is significantly curtailed. For example, a computer programme that has been written to automatically generate items from an encyclopaedia entry about the rules of cricket would not require the programmer to have any knowledge of cricket. In contrast, semantic approaches require test developers to have "a deeper understanding of the input" (Kurdi et al., 2020, p. 137) and generally require content experts to be involved in the item development process.

- The *procedure of transformation* dimension determines the process by which the items are automatically generated from the input information. Within this dimension, there are **template-based**, **rule-based**, and **statistical** approaches. The template-based approach requires test developers to think deeply about what they are assessing and how to do this. Each item generated via the template-based approach will subsequently be reviewed by a subject matter expert to ensure it is of a high quality. In a rule-based approach, items are automatically generated by applying various "rules" to the input information. In comparison, statistical approaches to AIG are still largely in the development stage and involve using machine learning tools and artificial intelligence to automatically generate items.

Approaches to AIG can be conceptualised as combining a syntactic or semantic level of understanding *with* a template, rule, or statistical procedure to transform the input material into items.

The Scope of this Research

For the purpose of this paper, a computer-aided search for relevant literature was completed using the Dublin City University Library Database. Once relevant articles were identified, the reference lists were then screened in order to find further studies of interest. An additional literature search was also conducted using Google Scholar. The remainder of this paper focuses primarily on methods of AIG that combine a semantic and template approach to generate multiple-choice items, as this represents the bulk of peer-reviewed work on AIG that has been conducted thus far. Additionally, these are addressed because they represent the approaches that are most relevant to the field of certification and licensure testing. As items included in large-scale assessments in this area are typically developed with significant input from content experts and undergo a rigorous process of evaluation, it is unlikely that the generation of items with no input from experts and with no one involved who has a deep understanding of the material will be used widely in the short term. For example, statistical approaches are an exciting prospect; however, large amounts of data/information are needed to train these models (von Davier, 2019). Additionally, at present, it may prove difficult to defend test results stemming from items that have limited human input in their development. It is therefore unlikely that this approach will be easily accessible for all assessment organisations/providers (Thompson, 2019) in the immediate future. The authors do, however, acknowledge that alternative approaches to AIG will become more prominent in the future as they are further developed.

Key Findings: The Benefits of Automatic Item Generation

This section of the paper outlines some of the benefits of AIG, focusing on those that are most frequently cited in the literature.

1. **AIG can create a large number of items.**
   - One of the main benefits of AIG is that it can be used to develop large banks of items much more quickly than would be possible if the items were generated manually. Owing to this, test security is enhanced as

the risk of item exposure is significantly reduced. In contrast, traditional item development is a more time-consuming process in which subject matter experts are required to write each item individually (Gierl & Lai, 2016b). As such, it is often the case that the demand for items exceeds the supply, particularly with computer-based testing (CBT) increasing the frequency of testing. In large-scale assessments, it is beneficial for items to be used as infrequently as possible, in order to minimise the risk of prospective test-takers coming into contact with items that have been leaked.

- AIG has been cited as a major step forward in addressing these problems, with authors publishing on the use of AIG in various contexts consistently reporting a high number of generated items (see Gierl et al., 2016; Embertson & Kingston, 2018). The efficiency of developing items via AIG compared to traditional methods is clearly evidenced by the higher item output within a shorter time frame (Gierl et al., 2012).
- Research suggests that to ensure this benefit is realised, test developers need to determine how to best create items that are likely to be of sufficient quality to be used for their intended purpose (Embertson & Kingston, 2018). For example, if AIG were used to generate thousands of items, but a large proportion of these items were of insufficient quality, developers would still end up spending time and money on quality control processes to filter out unsuitable items, thus negatively impacting the cost-benefit rationale for AIG.

2. AIG can be cost- and time-effective.

- Another significant benefit of AIG that is frequently mentioned in the literature is that it is likely to be more cost- and time-effective than creating items manually. When items are created manually, human test developers are required to map out the content areas that the items will assess and then create a series of items one-by-one. However, with AIG, the role of human content experts is different. Content experts will outline the knowledge, skills, and content to be tested by the items, which are then generated using computer software (Gierl et al., 2012).
- While both AIG and traditional methods of item generation entail work by human content experts, the actual time it takes to produce items one-by-one should be significantly reduced using AIG. However, it is also noted that AIG may not be cheaper than manual item development in every case. For example, the initial upfront costs of creating AIG item models can be significant, as they require more time and expertise on behalf of content experts than creating items one-by-one (Kosh et al., 2019). This, coupled with the costs of developing software to generate items from the item models, means that AIG is likely to be more cost-effective on a per-item basis when a large number of items are produced/required.
- Furthermore, the cost-benefit analysis of AIG is affected by the extent to which created items need to be screened for quality control purposes. It is therefore incumbent upon test developers to determine whether AIG is likely to be the more cost-effective approach for their specific item development needs (Kosh et al., 2019).

Key Findings: Critiques Arising from the Literature

While AIG may be appealing to testing organisations whose demand for test items exceeds their current ability to develop these items (Pugh et al., 2020), there are critiques arising from the literature regarding this method of item development that must be considered.

1. **Quality assurance processes are an essential element of AIG.**
   - One of the most prominent critiques in literature focusing on AIG is concerned with quality assurance. Previous research suggests that items generated via AIG may not be of the same quality compared to items generated using traditional methods (Pugh et al., 2020). It is theoretically possible that test developers could face legal action from test-takers, alleging that AIG items are not equivalent to manually generated items, thus calling into question the validity of inferences made on the basis of awarded scores. As such, it is imperative that test developers can point to evidence regarding the equivalence of AIG and manually generated items.
   - A further area of concern pertaining to quality assurance is the risk of the efficiency of AIG being undermined if the generated items continue to require traditional and time-consuming review/evaluation processes (Embertson & Kingston, 2018). Currently, all items generated via AIG must meet qualitative criteria regarding content and must also have satisfactory psychometric properties. As noted by Embertson and Kingston (2018), if this review process cannot be reduced, then AIG will not perform as efficiently as it could in meeting the demand for test items. However, recent studies have indicated that a more limited, less expensive review process for items generated via AIG may be possible. For example, Embertson and Kingston (2018) investigated "to what extent are the psychometric properties across item variants from the same family predictable?" (p. 113). In their study, few psychometric differences were found, with item discrimination and item difficulty comparable across AIG and operational items. The authors argued that these promising results provide evidence that supports a shortened review process.
   - As AIG continues to develop and be introduced into various testing programmes, it is crucial that time and resources are dedicated to the quality process. This will ensure that generated items meet required qualitative and psychometric standards.
2. **Existing approaches to AIG do not address best practice for distractor generation.**
   - There are concerns expressed in recent literature regarding the quality of AIG distractors, with Kosh (2021) suggesting that the transition of AIG from theoretical research to operational implementation has led to unexpected challenges, particularly in the area of distractor generation. Some previous research has acknowledged that existing approaches to AIG did not address the best methods/practices for generating answer choices (Gierl, Lai, et al., 2015). For example, an established three-step process for automatically generating items put forward by Gierl et al. (2012) models the stem and the key; however, it does not specifically model the distractors (Gierl & Lai, 2013).
   - Research also draws attention to the fact that it is unfeasible to quality check distractors generated via AIG (Kosh, 2021). Because AIG can be used to generate large numbers of items (often in the tens of thousands; Gierl, Lai, et al., 2015), it is impossible to check the distractors for each generated item. In contrast, items developed via traditional methods are individually reviewed by subject matter experts. As such, distractors generated via AIG may not blend together and may confuse test-takers by including answer choices that are the same, answer choices that appear as outliers, or answer choices that are in different formats.
   - Any benefits gained in terms of generating more items will come with costs in terms of the additional time and resources spent ensuring quality (Royal et al., 2018). At the time of writing, methodological innovations are currently being developed to alleviate the challenges of automatically generating

distractors (for example, see Kosh, 2021). However, only methodological details are currently being published, and these novel approaches have not yet been operationally or psychometrically tested.

3. **The "automatic" element of AIG is questionable.**
   - While called automatic item generation, Gierl and Lai (2016a) state that AIG is anything but automatic. This is because AIG places significant demand on software developers and subject matter experts. Furthermore, it has been argued that because many current AIG approaches involve significant human preparation, they end up as "fill-in-the-blanks approaches" (von Davier, 2019, p. 2).
   - Perhaps one of the most important considerations regarding AIG is that if the testing community decides to embrace it, there will likely be challenges involved in extending it into everyday practice (Royal et al., 2018). For example, content experts will require professional development, while item writers and subject matter experts will need to learn how to use highly complex software to do a job they were already doing themselves. It is not known how these individuals will respond to this or how they will view this new process (Royal et al., 2018).

Recommendations

Based on the cited literature, the following recommendations should be considered by certification/licensure testing organisations that are considering incorporating AIG into their item development processes:

1. **Carry out an evaluation of their current item generation methods.** The potential for AIG to be used in the development of large numbers of items is well recognised. However, as there are also limitations to using AIG, it is recommended that testing organisations conduct an extensive evaluation of their current item generation methods, including a cost-benefit analysis. This analysis should include a review of the cost of content experts, software generation, and quality assurance procedures. Only after such an evaluation can a move to AIG be justified, as organisations will be able to meaningfully compare AIG methods to those methods currently in use.

2. **Develop unique evaluation procedures for items generated via AIG.** Currently, items that are developed using AIG require a rigorous process of evaluation before they can be included in any large-scale certification/licensure assessment. It is recommended that testing organisations develop an evaluation procedure for AIG items. This evaluation should include qualitative appraisal by content experts and extensive empirical piloting within assessment administrations.

3. **Select the most appropriate approach to AIG.** As reviewed, there are numerous approaches to AIG, and these are likely to be better suited to particular subject areas. At present, the recommended approach that is most applicable to certification and licensure assessments is the semantic approach combined with the template-based approach. As this requires the creation of bespoke item generation software, it is recommended that testing organisations investigate and determine the associated cost of creating similar programmes. It is also recommended that an evaluation be carried out to determine the most relevant approach to AIG for each specific subject area.

4. **Monitor advancements in AIG.** It is critical that testing organisations stay aware of advancements in AIG approaches (such as statistical methods). Research on these emerging approaches is likely to become more common and may be viable for testing organisations in the future. Organisations are therefore recommended to keep abreast of published work in this area.

Highlighted Areas for Future Research

Research focusing on AIG has proliferated in recent years but remains at an early stage overall. Considering this, there are several important directions for future research, four of which are outlined below.

1. **Higher-Order Cognitive Domains.** To date, a significant amount of research has sought to determine whether AIG can be used to create large numbers of multiple-choice items (for example, Gierl & Lai, 2016). However, these item types generally assess skills on the lower end of Bloom's taxonomy, such as remember and understand (Bloom, 1956). It remains to be seen whether AIG can be used to assess higher-order domains such as apply, analyse, evaluate, and create.

2. **Automating the Production of Templates/Item Models.** As noted earlier in this paper, the most common method for producing AIG items is for human content experts to develop an item model and for an item generator to produce the items from this model (see Gierl et al., 2012). However, it is evident that this approach requires significant resources, which impedes the cost-effectiveness of using AIG (Embertson & Kingston, 2018). Further research could explore how/if the item models themselves could be automated, to further reduce the cost and time commitments associated with AIG.

3. **Testing AIG in Different Domains.** To date, a significant amount of research in the area of AIG has focused on its application and potential uses in medicine (Gierl et al., 2016) and education (Kurdi et al., 2020; Embertson & Kingston, 2018). In theory, AIG should be a usable approach in any subject where there is a need for the development of large numbers of items (particularly multiple-choice items). Indeed, some research has examined the possibility of using AIG in other areas, such as language acquisition (Arendasy et al., 2011) and music psychology (Harrison et al., 2017). However, given the relative extent of the research completed in education and medicine, further research is needed to determine the applicability of AIG in other fields and whether there are unique challenges that arise when AIG is applied elsewhere.

4. **Evaluation Procedures.** As it stands, items generated using AIG are usually evaluated qualitatively by human content experts and/or through an extensive empirical trial (for example, by placing pilot items within larger tests). This process remains time-consuming and expensive, thus limiting the cost-effectiveness of AIG (Kosh et al., 2019). Further research is required to determine whether the evaluation process for AIG items can be streamlined or automated.

Conclusions

The main goal of this paper is to provide an overview of the prominent literature and research focusing on the topic of AIG, concentrating particularly on its relevance to large-scale certification and licensure assessments. Specifically, this paper details critiques arising from the literature and highlights the main benefits that testing organisations can expect if they choose to incorporate AIG. It is evident that AIG is becoming more common and represents a growing research area (Kurdi et al., 2020). This is likely to continue in the future owing to the extensive benefits that are associated with this method of item development when compared to traditional methods. However, it is also important that testing organisations take note of the current limitations connected with this approach to item development. It is clear that further research is required in order to strengthen the field. Moving forward, it seems likely that as further research is undertaken, more testing organisations will lean towards this new and innovative approach to developing test items.

Reference List

Arendasy, M. E., Sommer, M., & Mayr, F. (2011). Using automatic item generation to simultaneously construct German and English versions of a word fluency test. *Journal of Cross-Cultural Psychology, 43*(3), 464–479. https://doi.org/10.1177/0022022110397360

Bloom, B. S. (1956). *Taxonomy of educational objectives, Handbook 1:* Cognitive domain. New York: David McKay.

Ch, D. R., & Saha, S. K. (2020). Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies, 13*(1), 14–25. https://doi.org/10.1109/TLT.2018.2889100

Embertson, S. E., & Kingston, N. M. (2018). Automatic item generation: A more efficient process for developing mathematics achievement items? *Journal of Educational Measurement, 55*(1), 112–131. https://doi.org/10.1111/jedm.12166

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing, 12*(3), 273–298. https://doi.org/10.1080/15305058.2011.635830

Gierl, M. J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice, 32*(3), 36-50. https://doi.org/10.1111/emip.12018

Gierl, M. J., & Lai, H. (2016a). Automatic item generation. In Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.), *Handbook of Test Development* (2nd ed., pp.410-429). Routledge.

Gierl, M. J., & Lai, H. (2016b). A process for reviewing and evaluating generated test items. *Educational Measurement: Issues and Practice, 35*(4), 6-20. https://uploads-ssl.webflow.com/605d05264268210d8056643a/616259c62ebb56c744449abd_emip%20AIG%20review%20gierl%20lai%202016.pdf

Gierl, M. J., Ball, M. M., Vele, V., & Lai, H. (2015). A method for generating nonverbal reasoning items using n-layer modelling. In Computer Assisted Assessment. In Ras E., & Joosten-ten Brinke D. (Eds.) *Computer Assisted Assessment. Research into E-Assessment. TEA 2015.* Communications in Computer and Information Science, vol 571. Springer, Cham. https://doi.org/10.1007/978-3-319-27704-2_2

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*(8), 757-765. https://doi.org/10.1111/j.1365-2923.2012.04289.x

Gierl, M. J., Lai, H., Hogan, J. B., & Matovinovic, D. (2015). A method for generating educational test items that are aligned to the common core state standards. *Journal for Applied Testing Technology, 16*(1), 1-18. http://www.jattjournal.com/index.php/atp/article/view/80234/62031

Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A., & De Champlain, A. (2016). Evaluating the psychometric characteristics of generated multiple-choice test items. *Applied Measurement in Education, 29*(3), 196-210. https://doi.org/10.1080/08957347.2016.1171768

Harrison, P. M. C., Collins, T., & Mullensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Sci Rep, 7*(1), 3618. https://doi.org/10.1038/s41598-017-03586-z

Kosh, A. E. (2021). Distractor suites: A method for developing answer choices in automatically generated multiple-choice items. *Journal of Applied Testing Technology, 22*(1), 12-24. Retrieved from http://www.jattjournal.com/index.php/atp/article/view/155880

Kosh, A. E., Simpson, M., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost-benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice, 38*(1), 48–53. https://doi.org/10.1111/emip.12237

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30,* 121–204. https://doi.org/10.1007/s40593-019-00186-y

Pugh, D., De Champlain, A., Gierl, M., Lai, H., & Touchie, C. (2020). Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Research and Practice in Technology Enhanced Learning, 15*(12), 1-13. https://doi.org/10.1186/s41039-020-00134-8

Royal, K. D., Hedgpeth, M., Jeon, T., & Colford, C. M. (2018, January 8). Automated item generation: The future of medical education assessment? *EMJ Innov. 2*(1):88-93. https://www.emjreviews.com/innovations/article/automated-item-generation-the-future-of-medical-education-assessment

Thompson, N. (2019, December 9). What is automated item generation? https://assess.com/what-is-automated-item-generation/

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika, 83*(4), 847-857. https://doi.org/10.1007/s11336-018-9608-y

von Davier, M. (2019). Training Optimus Prime, M.D.: Generating medical certification items by fine-tuning OpenAI's gpt2 transformer model. *ArXiv, abs/1908.08594.* https://doi.org/10.48550/arXiv.1908.08594