

Challenges and Horizons of Artificial Intelligence in Cognitive Assessment and Educational Test Development: A Constraint- Satisfaction Conceptual Model

Yaser Dadkhah^{1*}, Saeed Setayeshi², Alireza Moradi³, Mir Mohsen Pedram⁴

¹ Ph.D. Candidate in Cognitive Neuroscience (Artificial Intelligence), Institute for Cognitive Science Studies, Tehran, Iran

² Professor, Department of Nuclear Engineering, Faculty of Physics and Energy Engineering, Amirkabir University of Technology, Tehran, Iran

³ Professor, Department of Clinical Psychology, Faculty of Psychology and Education, Kharazmi University, Tehran, Iran

⁴ Associate Professor, Department of Electrical and Computer Engineering, Kharazmi University, Tehran, Iran

*Corresponding author: Yaser.dadkhah@hotmail.com

Abstract

Repeated cognitive and educational assessment is indispensable in longitudinal research, in the evaluation of instructional and rehabilitation programmes, and in high-stakes testing. Yet re-administering the same items inflates scores through practice effects, exposes item content, and progressively erodes the validity of the measurement. Parallel forms are the classical remedy, but their manual construction is slow and costly, so most cognitive instruments never acquire one. This paper argues that the bottleneck is not conceptual but procedural, and that it can be dissolved by treating item construction as a formal search problem. We present an analytical model in which the construction of an item for a rule-governed cognitive task is expressed as a constraint satisfaction problem defined over shape variables, latent weights and finite domains, and governed by five psychometrically motivated constraints: balance, difficulty, error-based distractors, non-repetition, and chain necessity. A generator combining backtracking search with arc-consistency propagation is described, together with an operational difficulty function that encodes working memory load, mapping load, and inferential depth. The distinguishing property of the approach is that item quality is guaranteed structurally at construction time rather than verified after the fact, and that every generated item carries a fully auditable derivation. We then examine four horizons this capability opens for educational assessment (dynamic item banking and adaptive testing, test security in high-stakes examinations, unbiased measurement of change in instructional interventions, and diagnostic formative assessment) and close with the technical, psychometric and ethical limits that a symbolic generator must respect before it can be trusted with consequential decisions.

Keywords: Automatic item generation, Constraint satisfaction, Parallel forms, Educational assessment, Symbolic artificial intelligence

1. Introduction

Cognitive assessment underpins a wide range of decisions in education: identifying gifted learners and those in need of special provision, evaluating the effectiveness of instructional programmes, and tracking cognitive development over time. Almost all of these purposes share one structural requirement, the same construct must be measured more than once in the same person. This requirement collides with a well-

documented psychometric phenomenon. When a test is re-administered, scores rise systematically for reasons that have nothing to do with genuine ability growth: familiarity with the task format, memory for specific items, and refinement of response strategies. Meta-analytic evidence indicates that this practice effect is substantial in reasoning and memory instruments and is not fully dissipated even after a year (Calamia et al., 2012).

The classical remedy is the parallel form: two or more versions of a test that measure a single construct with equivalent difficulty, reliability and structure (Lord and Novick, 1968). The remedy is old and uncontroversial, yet in practice very few cognitive instruments possess a normed parallel form. The reason is economic rather than theoretical. Building one by hand requires an expert to author every item, followed by pilot administration, item analysis, revision and re-norming, a cycle measured in years. A second pressure compounds the first: because a fixed form is used for a long period, its content leaks, and every subsequent administration inherits the resulting threat to validity (Gierl et al., 2022).

Automatic item generation (AIG) offers a way out. In this paradigm items are produced not one at a time by an expert but by an algorithm operating on an explicit model of item structure (Gierl and Haladyna, 2013). The field has grown rapidly; a systematic review of 93 studies identified cost reduction and a continuous supply of items as its two principal drivers in education (Kurdi et al., 2020). What remains contested is which family of algorithms should do the generating. Data-driven generative models excel where the rules governing the content are implicit and hard to extract, but they require large training corpora, offer no guarantee that a generated item is correct, and expose no rationale for the properties of what they produce (von Davier, 2018; Sommer and Arendasy, 2025).

This paper takes the opposite position for a specific and well-delimited class of tasks. Where the logic of an item is explicit, finite and formalisable (as it is in visuospatial quantitative reasoning tasks such as balance-scale items) the appropriate instrument is symbolic. We therefore formulate item construction as a constraint satisfaction problem (CSP), one of the foundational frameworks of symbolic artificial intelligence (Mackworth, 1977; Russell and Norvig, 2021), and examine what this formulation makes possible for educational measurement. The contribution is analytical rather than empirical: we present the model, the algorithm and the architecture, and we map the applications and the limits.

2. Theoretical Framework and Related Work

2.1. Two generations of automatic item generation

Two broad approaches are distinguishable in the AIG literature. The first is template-based: a parent item is selected and its surface elements are declared variable, so that substitution yields a family of structurally similar items. The approach is simple and productive, but its control over psychometric properties is indirect, because the template guarantees surface similarity rather than cognitive equivalence. The second approach, formulated by Embretson (1998) as the cognitive design system, begins from a theory of the task: the processing components that determine difficulty are first identified and validated, and items are then generated so that those components take specified values. Freund et al. (2008) showed that most of the difficulty variance of computer-generated figural matrix items can indeed be explained by the structural rules used to build them, and Matzen et al. (2010) demonstrated large-scale generation of Raven-like matrix problems with normed properties.

The transfer of these methods into operational testing has been driven largely by medical education, where Gierl et al. (2012) showed that algorithmically generated multiple-choice items can substitute for hand-written ones. Arendasy and Sommer (2012) extended the argument to high-stakes educational and occupational assessment, with the important qualification that this is admissible only when automated quality-control mechanisms are embedded in the generation process itself.

2.2. Symbolic versus sub-symbolic generation

A more recent wave applies neural and transformer architectures to item production. Von Davier (2018) generated personality items with long short-term memory networks, and a critical review by Sommer and Arendasy (2025) compares rule-based and transformer-based generation directly. Their conclusion is that data-driven generators dominate on output diversity while remaining weak on exactly the two properties that standardised assessment requires most: guaranteed correctness and interpretability. The choice between the two traditions is therefore not a matter of fashion but of problem structure, and turns on three criteria, the availability of explicit structural knowledge, the need for auditability, and the volume of training data that can realistically be obtained. Commercial intelligence tests fail the third criterion by construction, since their item pools are confidential.

This is not an argument against learning-based methods in cognitive science; their value in modelling complex, weakly structured signals is well established, as work on deep cellular automata for biological growth processes (Nikraveshtin and Setayeshi, 2024) and on deep architectures for EEG-based brain–computer interfaces (Anjerani et al., 2025) illustrates. The argument is narrower: when the generative rules of a task are already known exactly, recovering them from data is a detour that costs correctness guarantees and transparency.

2.3. The gap addressed

Three gaps emerge from this literature. First, AIG research has concentrated almost exclusively on figural matrices and academic content; the subtests of individually administered intelligence batteries have received little attention. Second, generation has been dominated by template substitution, and the explicit formulation of item construction as a constraint satisfaction problem — which permits simultaneous control of several properties with a mathematical guarantee — is rare. Third, discussion of educational applications tends to remain at the level of item supply, without asking what an interpretable generator makes possible that an opaque one does not. The model presented below is directed at these three gaps.

3. The Proposed Constraint-Based Model

The target task class is rule-governed visuospatial quantitative reasoning. Its canonical instance is the balance-scale item, in which one or more two-pan balances establish weight relations among geometric shapes and the examinee must select the option that restores equilibrium of an incomplete balance. Solving such an item requires perceptual encoding of the shapes, construction of a shape-to-weight mapping, maintenance of that mapping in working memory, and algebraic substitution to recover the unknown.

3.1. Formal statement

A constraint satisfaction problem is defined by the triple in Equation (1), where X is a finite set of variables, D assigns a finite domain to each variable, and C is a set of constraints restricting the admissible combinations of values. A solution is a complete assignment that violates no constraint (Russell and Norvig, 2021).

$$P = \langle X, D, C \rangle \quad (1)$$

The decisive property of the target task for this formulation is that item correctness reduces to a small system of linear equations: if every shape type is assigned an integer weight, the equilibrium of each balance is one equation and the correct option is the unique solution of the system. Equation (2) states the equilibrium condition for a complete balance with left pan L and right pan R .

$$\sum w(l_i) = \sum w(r_j), \quad l_i \in L, \quad r_j \in R \quad (2)$$

In the instantiation described here, X comprises the shape types used in the item, the latent integer weight of each type, the pan configurations of each balance, the keyed option and four distractors. The domains D are deliberately bounded (five shape types, integer weights from 1 to 6, and at most ten tokens per pan) which keeps perceptual load comparable to conventional items and keeps the search space finite and tractable.

3.2. The five constraints

Constraints are where psychometric theory enters the algorithm. Each of the five in Table 1 encodes a requirement that would otherwise be checked by a human reviewer after the item had been written.

Table 1. The five constraints, their operational definition and psychometric rationale

Constraint	Operational definition	Psychometric rationale
Balance	Every complete balance satisfies Equation (2)	Guarantees existence and uniqueness of the key
Difficulty	Index f of Equation (3) falls inside the band of the target level	Aligns the difficulty distribution with the reference form
Error-based distractors	Each incorrect option is the endpoint of a specific faulty solution path	Raises item discrimination and yields diagnostic information
Non-repetition	No item is isomorphic to an earlier item or to a reference item	Preserves diversity; prevents superficial re-skinning
Chain necessity	The key is reachable only by using every balance presented	Makes inferential depth a valid component of the difficulty index

Difficulty is not left to intuition. It is defined operationally in Equation (3) as a weighted sum of three structural quantities: n_shapes , the number of tokens displayed, standing for working-memory load; n_types , the number of distinct shape types, standing for the load of maintaining the shape-to-weight mapping; and n_steps , the number of substitutions required to reach the key, standing for inferential depth. The coefficient of 2 on the third term reflects the position of inferential depth as the dominant determinant of difficulty in quantitative reasoning.

$$f = n_shapes + n_types + 2 \times n_steps \quad (3)$$

A worked illustration makes the mechanism concrete. Consider a two-balance item in which one star equals two triangles and the known pan of the incomplete balance holds one star and one triangle. Five tokens are displayed, so $n_shapes = 5$; two distinct types appear, so $n_types = 2$; and one substitution is needed, so $n_steps = 1$. Equation (3) then yields $f = 5 + 2 + 2 = 9$. An identity relation contributes no inferential step and therefore produces the lowest attainable values of f . Because the generator rejects any candidate whose index falls outside the band of the requested level, the difficulty distribution of a generated form is fixed before generation begins rather than discovered afterwards. Figure 1 shows an item produced by the generator together with its five options.

$$f = 5 + 2 + 2 \times 1 = 9$$

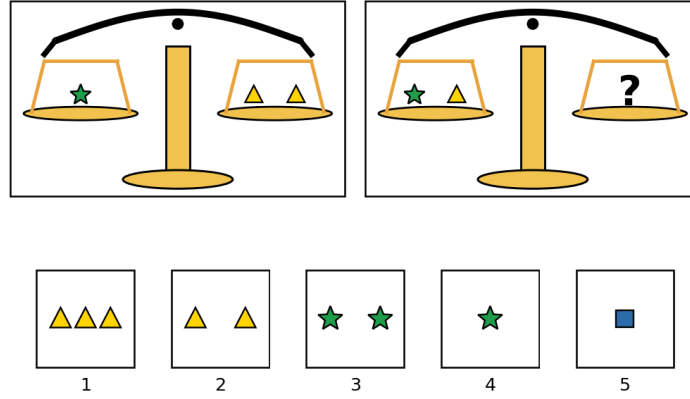


Figure 1. A generated balance-scale item with its five response options and the associated difficulty computation

3.3. Search strategy and generation algorithm

Solutions are recovered by backtracking search, a depth-first procedure that assigns values to variables one at a time and returns to the previous variable whenever the partial assignment becomes inconsistent. Its efficiency depends critically on constraint propagation. Arc consistency, implemented by AC-3, repeatedly examines every arc of the constraint network and removes from each domain those values for which no consistent partner exists, pruning large parts of the state space before the search reaches them (Mackworth, 1977). Algorithm 1 states the full procedure.

Algorithm 1. Constraint-based generation of a calibrated item set

Input: target difficulty level d , required number of items n
Output: set S of generated items, $|S| = n$

- 1: initialise shape and weight domains; $S \leftarrow \emptyset$
- 2: while $|S| < n$ do
- 3: sample structural parameters (tokens, types, balances) consistent with $\text{band}(d)$
- 4: assign latent weights to shape types by backtracking search under the balance constraint
- 5: enforce arc consistency (AC-3); prune inconsistent values from all domains
- 6: if a contradiction is reached then backtrack to step 4 and try the next assignment
- 7: compute the key; derive four distractors from the error typology
- 8: verify option distinctness, key uniqueness and non-isomorphism against S and the reference form
- 9: if every verification passes then $S \leftarrow S \cup \{\text{item}\}$ else discard and return to step 3
- 10: end while
- 11: return S

Two properties of Algorithm 1 deserve emphasis. First, the verification block in steps 8 and 9 is not an error-detection stage but a documentation stage: any candidate surviving the search has already satisfied every declared constraint by construction. Second, the output is a decomposable specification of the item (balance structure, latent weights, options, key and difficulty components) rather than a picture. Rendering is a separate downstream layer, which means the visual presentation can be redesigned without touching the generative logic.

3.4. System architecture

Figure 2 places the generator inside a complete assessment pipeline. The cognitive model of the task supplies the constraint set; the CSP engine produces specifications; a rendering layer converts them into stimuli; and a delivery platform administers them. Because generation is inexpensive, the engine may be invoked either offline, to populate a bank, or at run time, to construct a form that has never existed before the examinee sits down.

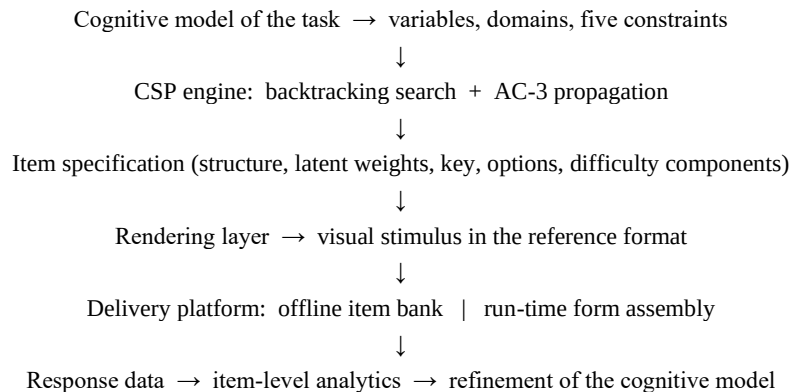


Figure 2. Schematic architecture of the constraint-based item generation pipeline. The feedback path from response data to the cognitive model closes the loop between measurement and modelling.

The engine was implemented in Python with the open-source python-constraint solver; item specifications are stored as JSON and rendered with matplotlib. In the instantiation used for illustration, the generator was configured to produce twenty-seven five-option items distributed evenly across three difficulty bands, matching the structure of a reference subtest.

4. Educational and Psychological Horizons

4.1. Dynamic item banks and adaptive testing

Computerised adaptive testing selects each successive item from the examinee's current ability estimate and consequently consumes items at a rate that manual authoring cannot sustain (van der Linden and Glas, 2000). A constraint-based generator can feed such a bank continuously. The gain is amplified by a finding that recurs across the AIG literature: when the generative model is strong, items from the same family are equivalent enough that psychometric parameters may be estimated at the family level rather than item by item (Gierl and Lai, 2016), which removes much of the pre-testing cost that makes large banks prohibitive.

4.2. Test security in high-stakes examinations

Item exposure is a permanent threat to national and entrance examinations. Gierl et al. (2022) showed that combining content coding with automatic generation improves the security of an item bank systematically rather than incidentally. The architecture of Figure 2 permits a stronger measure. Because a form costs seconds to generate, the generator can be embedded in the delivery platform so that forms are assembled at the moment of administration. Under this design there is no fixed form to leak, and exposure control ceases to be a matter of administrative vigilance and becomes a property of the system.

4.3. Valid measurement of instructional change

Evaluating an educational or rehabilitative intervention normally requires a pre-test–post-test design, and the internal validity of such a design rests entirely on separating genuine change from practice. When the second administration presents items that share the structure but not the content of the first, the memory

advantage disappears and the difference score becomes interpretable. This concern is not hypothetical in cognitive training research, where repeated measurement is intrinsic to the design (Salamat et al., 2024).

4.4. Formative assessment and cognitive diagnosis

Contemporary practice has shifted from assessment of learning to assessment for learning, which multiplies the demand for items still further. Here the error-based distractor constraint yields a benefit that item quantity alone does not. Because every incorrect option is the terminus of a defined faulty reasoning path (miscounting, copying a pan without inference, substituting the wrong type) the option a learner selects identifies the misconception rather than merely recording a failure. The generator thus produces diagnostic information as a by-product of its construction logic, which an opaque generator cannot supply.

4.5. Reducing the cost of localisation

For assessment systems outside the English-speaking world, norm obsolescence and the expense of re-norming are chronic constraints, and validated parallel forms are correspondingly scarce. Balance-scale items are entirely non-verbal and the constraints are defined over numerical relations among shapes rather than over language, so the generator transfers across linguistic contexts without translation or cultural adaptation. This makes the approach a plausible template for building local item banks for other rule-governed subtests.

5. Challenges, Ethical Considerations and Conclusion

5.1. Technical and psychometric limits

Three limits bound the claims made here. The first is scope. The approach presupposes a task whose correctness criterion is explicit and formalisable; it applies naturally to balance-scale, matrix and analogous reasoning tasks, and it does not transfer to vocabulary, comprehension or general-knowledge items whose correctness depends on linguistic and cultural knowledge that resists axiomatisation. The second is computational. Enlarging the domains (more balances, wider weight ranges) expands the state space exponentially, and without careful propagation the generator degrades from seconds to intractability. Arc consistency is what keeps the problem practical, not an optional refinement.

The third and most consequential limit is that structural guarantee is not psychometric equivalence. That an item satisfies every declared constraint is a statement about its construction, not about how examinees behave when they meet it. Two forms can be structurally matched and still differ in perceived difficulty. Empirical validation therefore remains obligatory before any generated form informs a consequential decision, and item-level response data analysed within an item response theory framework (Embretson and Reise, 2000) is the natural instrument for that validation.

5.2. Fairness, accountability and the role of the human expert

Automated test construction raises questions that are ethical before they are technical. Algorithmic fairness is the first: a generator must not encode structural advantage for any group, and the interpretability of a symbolic system is a material asset here, because each constraint is inspectable and each item auditable, which is not true of a generator whose rules are distributed across network weights. Accountability is the second. The algorithm functions as a construction assistant; responsibility for construct validity, standardisation and score interpretation remains with the psychometrician. Presenting a generator as a replacement for professional judgement rather than an instrument of it would misrepresent both what it does and what it can be held to.

5.3. Conclusion

This paper has argued that for rule-governed cognitive tasks the long-standing shortage of parallel forms is a procedural rather than a conceptual problem, and that expressing item construction as a constraint

satisfaction problem dissolves it. The model relocates quality assurance from post-hoc verification to structural guarantee at construction time, requires no training corpus, and leaves a complete audit trail for every item it produces. For educational measurement this opens a path towards dynamic item banks, forms assembled at run time, defensible measurement of instructional change, and distractors that diagnose rather than merely penalise.

The horizon is hybrid rather than symbolic. Constraint-based engines are the right instrument where the rules are known and auditability is required; learning-based models are the right instrument where structure must be inferred from data. A generation architecture that assigns each family to the problems it suits, symbolic control over psychometric constraints, statistical modelling over what remains implicit, is the more promising direction for the next generation of intelligent, fair and dynamic assessment tools.

References

- Anjerani, Marzieh, Pedram, Mir Mohsen, Mirzarezaee, Mitra and Malekian, Ehsan. (2025). Data augmentation and feature extraction using deep learning for motor imagery EEG-based brain computer interface classification. *Neural Computing and Applications*.
- Arendasy, Martin E. and Sommer, Markus. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences*. 22 (1). 112–117.
- Calamia, Matthew, Markon, Kristian and Tranel, Daniel. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*. 26 (4). 543–570.
- Embretson, Susan E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*. 3 (3). 380–396.
- Embretson, Susan E. and Reise, Steven P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Freund, Philipp A., Hofer, Stefan and Holling, Heinz. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*. 32 (3). 195–210.
- Gierl, Mark J. and Haladyna, Thomas M. (Eds.). (2013). *Automatic Item Generation: Theory and Practice*. New York: Routledge.
- Gierl, Mark J. and Lai, Hollis. (2016). Automatic item generation. In Lane, S., Raymond, M. R. and Haladyna, T. M. (Eds.), *Handbook of Test Development*. 2nd edition. New York: Routledge. 410–429.
- Gierl, Mark J., Lai, Hollis and Turner, Simon R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*. 46 (8). 757–765.
- Gierl, Mark J., Shin, Jinnie, Firoozi, Tahereh and Lai, Hollis. (2022). Using content coding and automatic item generation to improve test security. *Frontiers in Education*. 7. Article 853578.
- Kurdi, Ghader, Leo, Jared, Parsia, Bijan, Sattler, Uli and Al-Emari, Salam. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*. 30 (1). 121–204.
- Lord, Frederic M. and Novick, Melvin R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
- Mackworth, Alan K. (1977). Consistency in networks of relations. *Artificial Intelligence*. 8 (1). 99–118.
- Matzen, Laura E., Benz, Zachary O., Dixon, Kevin R., Posey, Jamie, Kroger, James K. and Speed, Ann E. (2010). Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods*. 42 (2). 525–541.
- Nikravesh Matin, Hossein and Setayeshi, Saeed. (2024). A computational tumor growth model experience based on molecular dynamics point of view using deep cellular automata. *Artificial Intelligence in Medicine*. 148. 102752.
- Russell, Stuart J. and Norvig, Peter. (2021). *Artificial Intelligence: A Modern Approach*. 4th edition. Hoboken, NJ: Pearson.

- Salamat, Marziyeh, Moradi, Alireza, Hasani, Jafar et al. (2024). Exploring the efficacy of memory specificity training on depression among Iranian adolescents: A comparative analysis of online vs. in-person delivery. Scientific Reports. 14. Article 22412.
- Sommer, Markus and Arendasy, Martin E. (2025). Automatic- and transformer-based automatic item generation: A critical review. Journal of Intelligence. 13 (8). Article 102.
- van der Linden, Wim J. and Glas, Cees A. W. (Eds.). (2000). Computerized Adaptive Testing: Theory and Practice. Dordrecht: Kluwer Academic Publishers.
- von Davier, Matthias. (2018). Automated item generation with recurrent neural networks. Psychometrika. 83 (4). 847–857.