Literature review

Introduction

Cognisess provide web based psychometric games that build a profile of an user's cognition, emotion, skills, social cohesion, behaviour, linguistics and health [1]. Over several years the company has collected gameplay data from hundreds of thousands of participants; constructing a valuable database. Cognitive assessments are designed in accordance with the Cattel-Horn-Carroll (CHC) theory of intelligence. This framework conceptualises intellect as a hierarchy of general, broad and narrow cognitive abilities [2]. Moreover, Cambridge Neuropsychological Test Automated Battery (CANTAB) offer validated computerised cognitive assessments that measure cognitive domains that can be mapped onto CHC constructs [3]. CANTAB, therefore, provides benchmarking for Cognisess for tests assessing CHC-linked constructs. This review aims to examine how modern psychometric models, such as item response theory (IRT) and its deep learning and Bayesian variants, can guide analysis of large-scale gameplay data and inform the design of computerised adaptive testing (CAT) systems. Such approaches could streamline game development and enhance the precision with which games discriminate between underlying cognitive abilities.

CHC theory and CANTAB

Dimensions of ability, according to CHC theory, have hierarchical structure. The lowest order of this hierarchy is referred to as specific abilities (for example, remembering numbers), which are the only assessable elements of this framework. The other orders of this theory are narrow, broad and general abilities. Narrow abilities are groups of highly correlated specific abilities. Next, sets of highly related narrow abilities make up broad abilities (eg: auditory short-term storage and visual short-term shortage – working

memory capacity). Broad abilities are denoted as *general,* G and the specific area is given by the subscript. Such that working memory capacity would be G_{wm}. Carrol submitted evidence in 1993 for the existence of 8-10 broad collections of narrow abilities, however the exact number is still an area of research [4].

CANTAB assessments were created in the 1980s display satisfactory discriminant abilities between healthy adults and cognitively clinical populations such as Alzheimer's or attention deficit hyperactivity disorder. However, when used beyond the clinical setting, CANTAB has been shown to show modest associations with traditional neuropsychological assessments [5].

Cognisess's game-based assessments enable the measurement of CHC-related abilities through more engaging and task-relevant tests. Several of these games are analogous to established CANTAB tasks, some of which have demonstrated validity in peer-reviewed research. Moreover, Cognisess's assessments capture performance across multiple cognitive domains and, due to their scalability, generate large volumes of item response data that can be mapped onto CHC-defined constructs.

IRT Foundations

IRT gives a mathematical basis for grasping the relationship between and individuals' latent ability and the probability that said individual will answer correctly in response to particular questions, or items. IRT makes assumptions that performance on a given item is explicable with a small number of parameters describing both the item and the person. The central parameters are item difficulty (b), item discrimination (a) and the persons latent ability (θ). Difficulty (b) is indicative of how challenging the item is; the higher the value of b, the higher the probability required for 50% chance of answering correctly.

Discrimination refers to how efficiently an item can discern players of different abilities. θ is typically modelled as a standard normal distribution on a continuous scale, such that a higher value indicates a greater ability with respect to this latent trait [6].

A restriction of unidimensional IRT is its assumption that all items correspond to a distinct underlying ability. However, cognitive assessments normally draw on multiple interwoven domains. Multidimensional item response theory (MIRT), as outlined by Recklase [7], furthers IRT to account for several latent traits simultaneously. MIRT influences each items response probability with a vector of latent abilities, as opposed to a single value of θ. MIRT is, therefore, structurally more compatible with theories of intelligence that conceptualise multiple, related abilities – such as CHC.

Bayesian IRT, as described by Fox [8], extends the IRT framework by adding hierarchical structures and prior information into parameter calculation. Bayesian IRT quantifies uncertainty and updates parameter estimates in line with new data becoming available. Thus, large-scale longitudinal analyses are fitting for this approach. Taken together, classical IRT, MIRT and Bayesian IRT are a logical set of approaches to give coherent and scalable modelling for latent traits in cognition-based gameplay data, offering statistically robust and explainable simulation tools.

Modern developments in IRT

Advances in psychometrics have extended IRT beyond traditional parametric models to exploit complex, high-dimensional data. Deep Learning–Enhanced IRT, developed by Cheng and Liu [9], integrates neural representation learning with the interpretability of classical IRT. Their Deep Item Response Theory (DIRT) framework employs deep neural

networks to derive item and person parameters from auxiliary information such as question text or tagged cognitive concepts. These inferred parameters—difficulty (b), discrimination (a), and proficiency (θ)—are then used within the IRT response function. This architecture allows DIRT to capture non-linear relationships between item features and latent ability while retaining the explanatory semantics of IRT. By leveraging embeddings and attention-based layers, DIRT can incorporate semantic and structural information from tasks, improving predictive performance in sparse or complex datasets [9], [10].

Building on this, recent studies have incorporated Bayesian estimation within deep IRT models as a practical extension of earlier Bayesian IRT principles [8]. In this context, Bayesian methods help stabilise the training of neural networks, manage overfitting through prior information, and provide measures of uncertainty for the estimated parameters. They also allow the model to update ability estimates as new data are collected, which is particularly useful for longitudinal or adaptive assessments such as gameplay analytics.

Further developments in dynamic and non-parametric IRT have expanded the framework's ability to capture continuous changes in ability over time. The Generalised Dynamic Gaussian Process IRT model proposed by Chen, Montgomery and Garnett [11] combine Gaussian process priors with ordinal response modelling, enabling smooth, time-evolving estimates of latent traits and flexible, non-linear item–response functions. Such methods are especially relevant to repeated or longitudinal cognitive gameplay data, where ability or engagement may fluctuate dynamically.

Together, these developments mark a shift toward more adaptive, data-rich, and explainable psychometric modelling. Deep IRT approaches, informed by Bayesian estimation, preserve interpretability while scaling to the complexity of digital assessment data, bridging the gap between traditional psychometrics and modern machine learning.

Extensions and Alternatives

Beyond documented forms of IRT, several related latent-variable models demonstrate how psychometric inference can proceed even without pre-defined correct responses. Test Theory Without an Answer Key**,** introduced by Batchelder and Romney [12], addresses situations in which the "true" answers are unknown to the researcher. Instead of scoring responses against an external key, the model estimates both respondents' competencies and the latent cultural consensus underlying their answers. This framework allows recovery of a shared "truth" within a coherent group, even from purely binary judgments. It is particularly relevant for unscored or exploratory game data, where participant behaviour reflects partial knowledge or implicit norms rather than discrete right–wrong outcomes [12].

Building on this foundation, Cultural Consensus Models further the idea through hierarchical Bayesian estimation [13]. Anders et al. (2014) generalised consensus modelling to continuous and ordinal responses, allowing for heterogeneous subcultures with distinct shared beliefs. The model jointly infers group-level "consensus truths," item difficulty, and individual bias or knowledge, providing a probabilistic structure for grouping individual judgements.

Together, these models illustrate that latent-variable inference is not restricted to standard test scoring or performance metrics. They demonstrate how consensus or competence can be derived from patterns of agreement, offering a principled basis for analysing open-ended or behavioural data from cognitive gameplay tasks where no explicit answer key exists.

Adaptive Testing and Future Directions

CAT represents a practical application of IRT, using its probabilistic backbone to tailor assessments to each individual's ability level. Rather than administering a fixed set of items, CAT algorithms dynamically select the next question that maximises information about a test-taker's latent trait (θ), typically choosing items whose difficulty (b) aligns with the individual's current ability estimate. This iterative process continues until the ability estimate reaches a specified level of precision, reducing test length while maintaining measurement accuracy [7], [8].

Recent evidence supports the effectiveness of IRT-based adaptive testing. Huda et al. [14] developed a web-based CAT platform implementing a three-parameter IRT model to adapt question difficulty and discrimination in real time. Tested with 90 students, the system achieved over 90% ratings for practicality and usefulness, suggesting that adaptive, IRT-driven testing improves both efficiency and engagement compared with static assessment formats. These findings demonstrate the feasibility of translating IRT theory into responsive, user-friendly assessment systems.

Extending this principle to game-based cognition assessment offers an opportunity for adaptive, data-driven personalisation. As gameplay data accumulate, the platform

could infer a player's ability distribution in real time and adjust the presented tasks accordingly—selecting levels, stimuli, or challenges that are neither too easy nor too difficult. Incorporating modern IRT extensions such as deep or Bayesian models could enhance adaptability by capturing non-linear patterns in behaviour, allowing ability estimates to update iteratively as players engage with diverse cognitive tasks.

If future analyses produce calibrated item parameters—each question's discrimination (a) and difficulty (b) values—these could serve as the foundation for adaptive design. A mapping from game design features to IRT parameters could be developed, either through straightforward correlations between task properties and a–b values, or via more complex deep learning methods that embed visual, textual, or structural features into a representational space. Such models could enable prediction of IRT parameters for unseen items, supporting scalable, automatic calibration across games.

Ultimately, integrating CHC-informed task design with adaptive deep and Bayesian IRT methods provides a clear route toward scalable, interpretable, and personalised cognitive game assessment—where each interaction could dynamically refine both measurement precision and player experience.

References

[1] Yondur, "Assessments," *Yondur.com*, 2025. [Online].
Available: https://app.yondur.com/assessments. [Accessed: 23-Oct-2025].

[2] W. J. Schneider, "The Cattell–Horn–Carroll theory of cognitive abilities,"
in *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, 4th ed., D. P. Flanagan and E. M. McDonough, Eds. New York, NY: The Guilford Press, 2018, pp. 163–202.

[3] Cambridge Cognition, “Digital cognitive assessments,” *Cambridge Cognition Ltd.*, 2025. [Online]. Available: https://cambridgecognition.com/digital-cognitive-assessments/. [Accessed: 23-Oct-2025].

[4] J. B. Carroll, *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge, U.K.: Cambridge University Press, 1993.

[5] P. J. Smith, A. C. Need, E. T. Cirulli, O. Chiba-Falek, and D. K. Attix, “A comparison of the Cambridge Automated Neuropsychological Test Battery (CANTAB) with ‘traditional’ neuropsychological testing instruments,” *Journal of Clinical and Experimental Neuropsychology*, vol. 35, no. 3, pp. 319–328, 2013, doi: 10.1080/13803395.2013.771618.

[6] F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum, 1980

[7] M. D. Reckase, *Multidimensional Item Response Theory*. New York, NY: Springer, 2009.

[8] J.-P. Fox, *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer, 2010.

[9] S. Cheng and Q. Liu, “DIRT: Deep Learning Enhanced Item Response Theory for Cognitive Diagnosis,” *Proc. 28th ACM Int. Conf. on Information and Knowledge Management (CIKM)*, 2019, pp. 2513–2521.

[10] Z. Song, S. Cheng, and Q. Liu, “Deep Learning Enhanced Cognitive Diagnosis for Intelligent Education Systems,” *arXiv preprint arXiv:1904.11738*, 2019.

[11] Y. Chen, J. M. Montgomery, and R. Garnett, “A Dynamic, Ordinal Gaussian Process Item Response Theoretic Model,” *arXiv preprint arXiv:2504.02643*, 2025.

[12] W. H. Batchelder and A. K. Romney, “Test theory without an answer key,” *Psychometrika*, vol. 53, no. 1, pp. 71–92, 1988.

[13] R. Anders, Z. Oravecz, and W. H. Batchelder, “Cultural consensus theory for continuous responses: A latent appraisal model for information pooling,” *Journal of Mathematical Psychology*, vol. 61, pp. 1–13, 2014.

[14] A. Huda, F. Firdaus, D. Irfan, Y. Hendriyani, A. Almasri, and M. Sukmawati, “Optimizing Educational Assessment: The Practicality of Computer Adaptive Testing (CAT) with an Item Response Theory (IRT) Approach,” *Int. J. Inform. Vis.*, vol. 8, no. 1, pp. 473–480, 2024.