The Rise of Immersive Cognitive Assessments: Towards Simulation-Based Assessment for Evaluating Applicants

Rebecca Kantar[1], Keith McNulty[2], Erica L. Snow[1](✉), Richard Wainess[1], Sonia D. Doshi[1], Devon B. Walker[1], and Matthew A. Emery[1]

[1] Imbellus, Los Angeles, CA, USA
esnow@imbellus.com
[2] McKinsey & Company, London, UK

Abstract. Imbellus is an assessment company that aims to test cognitive processes within the context of immersive simulation-based assessments. This paper explores our work with McKinsey & Company, a best-in-class management consulting firm, to build a simulation-based assessment that gauges applicants' cognitive skills and abilities. Leveraging a cognitive task analysis grounded in theoretical work and practical observations of on the job activities, we defined key work activities and skills needed to complete them. We then developed scenarios that abstracted and generalized the most crucial skills. To make sense of significant telemetry data from users' interactions with the assessment, we applied theoretically grounded expert models to guide our scoring algorithms. Our assessment draws inferences across seven major problem-solving constructs. We will present our initial findings and describe implications of our current work for the fields of artificial intelligence and assessment.

Keywords: Simulation-based assessment · Cognitive skills
Artificial Intelligence

1 Introduction

Imbellus is an assessment company that aims to evaluate cognitive processes within the context of simulation-based assessments. We will deploy these assessments across a variety of industries, domains, and organizations. We have partnered with McKinsey & Company, a best-in-class management consulting firm, to gauge incoming applicants' cognitive skills and abilities.

The Imbellus assessment focuses explicitly on incoming applicants' problem-solving skills and abilities. We define problem solving as a cognitive process directed at achieving a goal when no solution is obvious to the user (Mayer 2014). In partnership with McKinsey & Company, we conducted a cognitive task analysis (Schraagen et al. 2000) to conceptualize how successful problem-solving abilities manifest in the workplace. We developed our understanding of problem-solving skills from on-site interviews, case study analyses, and a review of related literature and

C. Penstein Rosé et al. (Eds.): AIED 2018, LNAI 10948, pp. 438–441, 2018.
https://doi.org/10.1007/978-3-319-93846-2_82

created a problem-solving ontology representing seven major constructs (e.g. situational awareness, metacognition, decision-making). We examined the structural alignment between our problem-solving ontology and the nature of employees' work by comparing job activities at McKinsey & Company. We mapped job activities to constructs to lay the blueprint for developing scenarios within our simulation-based assessment.

Our scenarios are tasks embedded within our assessment that abstract the context of a given work environment while maintaining opportunities for users to portray problem-solving capabilities required by the job. Transposing skills and applications to a different but comparable context allows us to assess far transfer (Perkins and Salomon 1992). Each scenario in our assessment is designed based on a set of problem-solving constructs and workplace activities. The assessment requires users to interact with a series of challenges involving terrain, plants, and wildlife within a natural world setting. This setting limits bias and offers an accessible context regardless of background and prior knowledge. For example, in one scenario a user may be identifying impending environmental threats in an ecosystem, given evidence. As a user interacts with our assessment, we collect a wealth of information about *how* they approach the task. Analyzing users' telemetry data (e.g. mouse movements, clicks, choices, timestamps), we can examine their cognitive processes and overall performance.

2 Overview of Score Development

From our research-driven, theoretical framework, we devised Imbellus scores to quantify *how* users' actions, timestamps, and performance within each scenario related to various cognitive constructs. Cognitive science, educational psychology, and learning science theories guided the mapping of each score to relevant constructs. Our scores focus both on the product (i.e., right or wrong) and on the process (i.e., how did they get there, what choices did they make, how many mistakes did they correct), which is more nuanced than traditional cognitive assessments.

We built, tested, and iterated upon Imbellus scores using both our theoretical framework and user data. We began our score design process by outlining an expert model, informed by our literature review of problem-solving skills, for each scenario. Our expert models outlined an expert's expected telemetry stream and corresponding evidence (e.g. efficiency, systematicity) for each assessment scenario. Expert models drove our evidence statements, outlining what information we would need to see from the user in the environment to infer strong problem-solving skills. For instance, if we wanted to measure informed decision making in our tasks, we would create an evidence statement that would define what informed decision making is and how it would manifest in our assessment. All scores were programmed using these evidence statements as the scoring parameters. After scores were built, we conducted think aloud testing and internal playtests to evaluate and iterate upon our initial expert models and scoring parameters. These initial scoring parameters served as the basis for our pilot study in November 2017.

3 Preliminary Pilot Overview

Using the preliminary Imbellus scores, we conducted a large-scale pilot study with McKinsey & Company. The goal of this pilot was to test our assessment platform and three scenarios and to examine the predictive power of our initial Imbellus scores. Information from this pilot study is being used to iterate and design future Imbellus scores and simulations.

3.1 Method

The pilot test was conducted in London, United Kingdom from November 13th to 17th, 2017. The test assessed 527 McKinsey candidates, of whom 40% were female, 59% were male, and 1% did not provide gender details. Of the pilot population, 56% of the participants were native English speakers, 43% were non-native, but fluent English speakers, and 1% had a business-level proficiency in English. The ethnic breakdown of the sample based on the Equal Employment Opportunity Commission (EEOC) guidelines was 52.6% White, 29.7% Asian, 3.9% Hispanic, 4.1% Mixed, 3.3% Black, 2.8% Other, and 3.5% who did not specify ("Code of Federal Regulations Title 29 - Labor" 1980).

The pilot test was an opt-in, proctored assessment following the participants' completion of the McKinsey & Company paper-based Problem-Solving Test (PST). The PST was validated using industry standard validation procedures for relevance to job specifications, scaling, and reliability. The Imbellus assessment was administered for 1 h. Over the 5 days of testing, a total of 29 testing sessions took place on McKinsey-owned laptops in an enclosed, proctored conference room setting. Following completion of the assessment, participants completed an online survey. The survey collected demographic information and feedback on each scenario's design and usability through 4-point Likert scales supplemented by open-ended questions.

3.2 Initial Results

Our scoring pipeline transformed each users' telemetry data into the Imbellus scores. To examine how well our assessment performed, we validated it against the PST. If the Imbellus scores are valid, we would expect a positive correlation between our scores and the PST. A PST score above a certain threshold is used as an early screen for cognitive ability in the McKinsey hiring process. The PST is one aspect of the McKinsey and Company selection process and is combined with other inputs. As a cognitive, work sample test the PST is likely to be a reasonable predictor of job performance. When the first cohort of applicants reaches their first performance review, they will be reassessed using job performance as the target. We built an elastic net logistic regression model trained on Imbellus scores to predict whether a user reached the PST threshold. We chose to use elastic net regularization because it tends to set the weights of uninformative scores to 0 while grouping predictive but near collinear scores (Hastie et al. 2009). We withheld 25% of the data for a test set.

The PST is a challenging test and fewer people reached the threshold than did not. This class imbalance means that models that predict mostly negative outcomes for everyone could have high accuracy. We assessed the model using the F_1-score, which is robust to class imbalances. The F_1-score is the harmonic average of precision (true positives divided by predicted positives) and recall (true positives divided by all positives). An F_1-score of 1 means the model is a perfect classifier; a model with an F_1-score of 0 is always wrong. The micro-averaged F_1-score for the test fraction was 0.621. This suggests that the Imbellus scores do have some predictive capability of users' cognitive skills but do not duplicate PST results.

Survey results indicated that 67% of users preferred the Imbellus assessment to the PST, and 91% of users found the Imbellus assessment engaging. Similarly, 64% reported the Imbellus assessment leveraged the same type of cognitive skills required for success in the McKinsey & Company selection process. These results suggest our assessment offers a more immersive alternative to existing assessment methods while maintaining context and construct validity.

4 Next Steps

Our forthcoming assessments are being designed for remote deployment via timed releases where users participate across any number of locations. To ensure no two assessments are the same, we are employing artificial intelligence (AI) approaches to scenario generation. We vary data-driven properties referenced across scenarios that, in turn, build unique versions of those scenarios. Our AI and data-driven architecture will protect against cheating and gaming of the test - a significant challenge facing many existing cognitive tests.

We are currently conducting playtests with McKinsey & Company employees and candidates globally while refining our assessment in preparation for operationalization next year. We are also developing additional ontologies and assessments for hard-to-measure skills and abilities. Our goal is to provide more specific, useful data on incoming applicants and employees that can inform the structuring of teams, assigning work, and managing talent.

References

Code of Federal Regulations Title 29 – Labor (1980). https://www.gpo.gov/fdsys/pkg/CFR-2016-title29-vol4/xml/CFR-2016-title29-vol4-part1606.xml

Mayer, R.E.: What problem solvers know cognitive readiness for adaptive problem solving. In: O'Neil, H.F., Perez, R.S., Baker, E.L. (eds.) Teaching and Measuring Cognitive Readiness, pp. 149–160. Springer, Boston (2014). https://doi.org/10.1007/978-1-4614-7579-8_8

Mayer, R.E., Wittrock, M.C.: Problem solving transfer. In: Berliner, D.C., Calfee, R.C. (eds.) Handbook of Educational Psychology, pp. 47–62. Simon & Schuster Macmillan, New York (1996)

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7

Schraagen, J.M., Chipman, S.F., Shalin, V.L. (eds.): Cognitive Task Analysis. Psychology Press, New York (2000)

Perkins, D.N., Salomon, G.: Transfer of learning. Int. Encycl. Educ. **2**, 6452–6457 (1992)