APPLIED PSYCHOLOGY IN HUNGARY

Special Issue

2023/3
VOLUME XXV

APPLIED PSYCHOLOGY IN HUNGARY

2023/3
Volume XXV

Year of Foundation: 1998

Published by the cooperation of Budapest University of Technology and Economics, ELTE Eötvös Loránd University and University of Debrecen, funded by the Hungarian Scientific Academy.

Address of the Editorial Office

Institute of Psychology, ELTE Eötvös Loránd University (ELTE PPK)
46 Izabella Street, Budapest, 1064

Editorial work and prepress

ELTE Eötvös University Press
E-mail: info@eotvoskiado.hu

Published by

the Dean of ELTE PPK

ISSN 3004-0906 (Online)
DOI: 10.17627/ALKPSZICH.2023.3

CONTENT

THE CONCEPT OF AN AI-BASED EXPERT SYSTEM, ATOM, FOR PREDICTING JOB SUCCESS – SPECIAL ISSUE –

THE CONCEPT OF AN AI-BASED EXPERT SYSTEM (ATOM) FOR PREDICTING JOB SUCCESS

Lajos Izsó
Budapest University of Technology and Economics
izso.lajos@gtk.bme.hu

Summary

Background and Aims: This article offers a brief prospective exposition of the other articles published in this special issue on artificial intelligence (AI)-based expert systems for predicting job success in general. It also focuses on a concrete implementation of such a system, developed by us. Apart from providing a broader perspective on the possibilities and limitations of applying AI in various fields of human resources management (HRM), such as recruitment, selection, employment, training, performance monitoring/management, wages, labor relations, occupational safety and health (OSH), etc., this text serves as the foreword by the editor of this special issue.

Recent AI applications supporting HRM are predominantly limited to recruitment; in other fields, companies still rely on traditional methods. Given the observed underperformance of HR personnel in workforce selection decisions, we have developed an AI-based system primarily for this task, known as ATOM (Artificial Intelligence for Testing Occupational success of Manpower). ATOM not only supports workforce selection but also effectively aids in recruiting and, to some extent, OSH.

The core function of ATOM is to “learn” the relationship between suitable predictors (variables suitable for predicting the future job success of applicants) and relevant success criteria scores for the given job. A notable feature of ATOM that provides outstanding efficiency and flexibility is its use of multiple machine learning (ML) algorithms running concurrently, with the results of the best-performing algorithm being accepted.

Keywords: artificial intelligence (AI), machine learning (ML), human resources management (HRM), job success, ATOM

DOI: 10.17627/ALKPSZICH.2023.3.5

Applying AI for supporting workforce selection: The state of affairs

When a new technology, like AI, starts getting noticed, subsequent hype is almost inevitable. As Kordon (2020) puts it, although "data is the new oil", it is time already in certain areas for the transfer from hype to real competitive advantage. Such a main area nowadays is, among many others, companies' staffing.

From a wider perspective, the proper handling of HR (human resources) at companies and other working organizations is of decisive importance. Namely, the HRM (human resources management) covers the main fields of recruiting, selection, employment, training, performance monitoring/ management, waging, labor relations, and occupational safety and health (OSH). While *recruitment* refers to the process of searching for potential applicants (and encouraging them to apply for an actual job), *selection* is the process of finding the best candidates from the shortlist created by appropriate preliminary testing. Selection is a decision-making process, which is still made mainly by humans, but appropriate artificial intelligence (AI) applications have tremendously high, yet far unutilized potential.

AI applications recently are being used to support HRM, still mainly in the field of recruitment. After Wheeler and Buckley (2021), it can be stated that "[c]ompanies still use traditional recruiting methods like job fairs, college recruiting, newspaper ads or billboards, and referrals; but the availability of data from multiple sources allows companies to proactively seek applicants who they then recruit to apply for jobs" (p. 60). Wheeler and Buckley (2021) later continue*,* "[s]ocial media sites like Facebook and LinkedIn allow digital recruiters to hyper-target possible applicants based upon the very information that users of those sites self-disclose" (p. 62).

Experience shows that humans, like HR persons, usually underperform in workforce selection decisions. We agree with Eubanks (2022), who states, "[a]dmit it: we're bad at selection. The data shows that the common ways we interview and many of the methods companies use to rank candidates (school attended, college grades, or other demographic data) are highly unreliable statistically. Translation: they are terrible as a gauge for whether someone can do a job or not" (p. 109). This is a strong argument for developing AI applications to support workforce selection. In addition to the advantage of possible better statistical reliability, such AI-based systems work incomparably quicker than humans do, and consequently, potentially much cheaper too.

The traditional (not AI-supported) workforce selection methods have serious validity limits. Barrick et al. (2001) state that even the personality construct of the best predictive validity – *"conscientiousness"* – usually has only a correlation of 0.20–0.25 with job performance.

Concerning the selection process two main biases are distinguished usually.

The first originates from the applicants, who are willing to show themselves as better than they are. This tendency might result in faked personality inventories and intentional fraud causing misinterpretation of resumés by HR coworkers, as Henle et al. (2019) published. The same distortions appear also in Assessment Centres, McFarland et al. (2003).

The second lies in the applied methods – König and Langer (2022) – since most personnel selection methods involve human decisions that are inherently error-prone.

In our opinion, however, there also exists a third bias. The source of this bias relates to the question: "Have we chosen the best procedure available in terms of given output-input relationships?"

The first bias remains henceforward in the cases of AI-supported methods also, while the second one, at least in principle, can be reduced by applying appropriate AI-driven methods. Reducing the third bias is only possible if a proper variety of procedures are used, either sequentially or simultaneously, and the results of the best performing one are accepted. Although this approach requires increased resources, it is already quite feasible for AI-driven methods running on today's quick computers. Notwithstanding, we have not found in the literature AI-based methods operating on this principle. Our system, however, to be reviewed in the next section, is based on this principle: it runs simultaneously many machine learning (ML) algorithms, and the outputs of the "winner" (the best performing one) are considered as final results. Thus, our system can effectively reduce this third type of distortion.

The rationale and functional fundamentals of ATOM

While the primary goal of psychology, as a scientific discipline, is to understand and explain human behavior, the practical fields of applied psychology – among them also work and organizational psychology – are much more interested in prediction (Yarkoni & Westfal, 2017). The explanation of the processes is usually not the goal of work and organizational psychology. Instead, its focus is usually on practical decision-making. ML methods can maximize the prediction accuracy of the models. While doing so, most of the time they do not provide an understandable explanation for how the phenomenon works. In this case, although it may provide a precise prediction for the given phenomenon, we will usually not know which variables played a role in the outcome and to what extent. This is the reason why we launched a project to develop a job success prediction system, for clearly practical purposes, based on AI and ML algorithms.

Of the HRM fields mentioned in the preceding section, our system, ATOM (Artificial intelligence for Testing Occupational success of Manpower) can mainly support recruiting and selection, and partly also OSH (refer to Izsó, Berényi & Pusker, this special issue). This special issue focuses on workforce selection by ATOM.

ATOM, developed by us at CIVIL Plc, is an AI-based expert system working on the web platform. The basic function of ATOM is to "learn" the relationship between suitable *predictors* and relevant *success criteria* of a certain job.

A predictor in this context is a variable suitable to predict the future job success of applicants.

Predictors can typically include, among many others: qualifications, relevant work experience, job-specific skills (e.g., driving license, computer proficiency, ability to speak a particular language), certain test scores, objective parameters measured by electromechanical or computerized aptitude-testing devices or work simulators, etc.

The job *success criteria* can typically be:

- actual quantitative and/or qualitative production data (however, such data – for theoretical or practical reasons – are not available for many jobs),

- management's scores on the employee's performance (the disadvantage of these is that they are generally not statistically reliable enough, primarily due to the so-called "halo effect" and "leniency" and "severity" biases).

The definition of well-founded success criteria should normally be an integral part of the job analysis.

Job analysis means the systematic collection and organization of information about the specific requirements (criteria) of the given work task for the employee. Therefore, it is desirable to compile so-called "job profiles" that contain these criteria in a well-structured way. If such criteria are available and appropriate – strongly correlated – predictors can also be found for them. Based on these predictors, the person's success in the given job can be predicted with a high probability.

While compliance of a candidate with the criteria can only be established later during the actual work activity, the predictors can be determined or measured by an instrument or simulator even before employment with relatively simple tools and at low cost.

ATOM can be applied if valid predictors are available for at least 100 employees already working in the given job, whose job success, varying from failure through medium to excellent success, is also available.

Then the ML algorithms in the core of ATOM can "learn" the relationship between the predictors (as input variables) and the criteria of job success (output variables). Based on the model built this way, ATOM later can predict the expected job success of new candidates from the predictors only, with a high probability.

A novel feature of ATOM is – as mentioned above and described in the 2nd article of this special issue (Gergely & Takács: this special issue) in more detail – that in its core many machine learning (ML) algorithms run concurrently, and the results of the best performing algorithm are accepted.

As can be read also in (Gergely & Takács, this special issue), the core of ATOM works via the type "supervised learning" of the ML, where the "training example" is a set of input-output data pairs. The goal of the process is classification, that is, to estimate probabilities for each new candidate falling into different success categories and then based on these, to determine success categories themselves solely from the predictors.

In the last, 6th article in this special issue (Izsó, Berényi & Takács, this special issue) real-life case studies are shortly presented. Here, applying the proposal of Tasdemir (2015), ROC analysis is used for evaluating ATOM's classification performance.

The main services of ATOM

ATOM package, corresponding to the three main user types, has three functionalities (sets of functions).

The employees' functionalities, by the help of which new candidates for a given job (or in the case of organization development, employees already working in the organization) can fill in the designated questionnaires or administer some simple instruments to themselves. The data collected this way can be processed from several points of view and in several directions. Among others, candidates – based on an automatic evaluation by ATOM – are provided with personal feedback about their strengths and competence fields still to be developed, and in case of interest, about jobs realistically available for them.

The employers' functionalities support the employers' HR and other co-workers in manpower selections (and later also in employees' career orientation and monitoring) by providing them with candidates' success probabilities for each success category.

The experts' functionalities support analyst experts – independent of the employer company –, with sophisticated analysis interaction possibilities and detailed feedback on the key competencies which have a significant impact on the success in the given job.

The sophisticated AI functions of ATOM, described in the previous section, support mainly interactions through the latest two (the employer's and the experts') functionalities.

The traditional process of workforce selection and supporting this process by ATOM

The traditional workforce selection with the help of work psychologists and/or occupational safety and health (OSH) professionals

The selection process is based on the following two kinds of expertise:

- *The expertise* (including relevant tacit knowledge) existing within the organization concerning the content of the given job, performance criteria, typical local conditions, and problems.
- *The expertise of work psychologists and/or OSH professionals* concerning human features and competencies (personality traits, work physiological characteristics, possibilities, and limits, etc.) and their assessment methods.

Both kinds of expertise – complementary to each other – are necessary. However, these two parties could only acquire their missing knowledge by investing a quite significant effort, and rather slowly and costly. Although there are examples that specific organizations (such as armed forces, nuclear power plants, airline companies, etc.) employ full-time psychologists who function in the organization and therefore have more profound knowledge concerning critical jobs, these are just exceptions.

Typical steps of traditional workforce selection:

- the organization invites work psychologists, OSH professionals, or a vocational advisor company who try to learn the work content and performance criteria;
- the invited experts assign competencies to the given job and also assign assessment methods (usually psychological tests or aptitude tests assessed by measuring devices) to these competencies;
- the organization (usually represented by its HR co-workers) together with the invited experts, organizes and executes the assessments (usually psychological testing or aptitude tests assessed by measuring devices);
- the invited experts compile personal expert reports;
- HR co-workers try to utilize these expert reports in their employment decisions.

All these steps are relatively slow, work-intensive, and costly. Slowness is especially problematic since it often happens that by the time the organization informs candidates about the results and employment decision, they are already employed by another company.

The workforce selection process supported by ATOM

Proposed steps of manpower selection supported by ATOM:

- the organization invites a company that possesses an ATOM package and also ATOM experts (ATOM experts have to learn the work requirements only very broadly);
- the invited ATOM experts apply a broad-spectrum general personality test, and therefore there is no need to select measuring instruments (except if unique competencies have to be assessed by particular questionnaires or measuring devices);
- the HR co-workers, together with the invited ATOM experts, organize and execute the assessments (if it is done online, it can be speedy);
- ATOM automatically – if only questionnaires are used – generates and forwards reports both to candidates and HR co-workers in real-time;
- HR co-workers utilize these reports in their employment decisions.

The process is mainly automated. Therefore, it is much less work-intensive, less expensive, and much faster.

Comparing traditional and ATOM-based manpower selection

The ATOM-based process also utilizes the two kinds of expertise mentioned above, but it is much simpler, quicker, and more reliable, since

- there is no need for organization-specific expertise other than the actual degree of job success for the employees included in the “learning sample”;
- and the expertise of vocational psychologists/physicians is utilized automatically via the algorithms of ATOM.

Potential possibilities of ATOM for applying in areas other than predicting job success

Although ATOM was developed as a sophisticated tool for predicting job success, under certain conditions, ATOM can also be applied in other areas. If we choose another goal function instead of job success, and we select predictors that are appropriate to this very other goal function, ATOM will, of course, produce predictions of the same accuracy as in the case of predicting job success. In short, it can be stated that ATOM is applicable to any prediction problems isomorphic with the predictors → job success schema.

The following *Table 1.* shows several examples from the possible many – from very different areas – for predictions that can be carried out by ATOM, and in which not job success is the goal function.

This table is to be interpreted in the following way: the goal function is a purposefully operationalized categorical measure of whether the examined event will occur within a predetermined period. The predictors are variables that influence these occurrences.

Table 1. Examples of prediction problems that ATOM can address with goal functions other than job success

Applications areas (Goal functions)	**Predictors** (Characteristics of the individuals that influence these occurrences)
work motivation of handicapped people (intention to return to work)*	marital status, place of residence *(farmhouse, village, small town, big city)*, living conditions *(without conveniences, with some conveniences, with all conveniences)*, salary, total income, living in family *(yes, no)*, if living in family *(number of family members living together, number of earning family members living together)*, years being unemployed, education *(no finished education, elementary school, secondary school, college, university, PhD)*, age, gender etc.
employees' turnover	salary, total income, place of residence *(farmhouse, village, small town, big city)*, living conditions *(without conveniences, with some conveniences, with all conveniences)*, living in family *(yes, no)*, if living in family *(number of family members living together, number of earning family members living together)*, education *(no finished education, elementary school, secondary school, college, university, PhD)*, age, gender etc.
churn, attrition *(cancelling e.g., telephone, cable TV, insurance, journal, etc. subscriptions by customers)*	financial situation *(heavy debt, moderate debt, no debt, properties)*, salary, total income, place of residence *(farmhouse, village, small town, big city)*, education *(no finished education, elementary school, secondary school, college, university, PhD)*, age, gender etc.
default in payment *(a borrower stops making the required payments on debt to banks or insurance companies)*	financial situation *(heavy debt, moderate debt, no debt, properties)*, salary, total income, place of residence *(farmhouse, village, small town, big city)*, education *(no finished education, elementary school, secondary school, college, university, PhD)*, age, gender etc.
voting to a particular political party	to which party voted earlier, participation in public life, religiousness, place of residence *(farmhouse, village, small town, big city)*, education *(no finished education, elementary school, secondary school, college, university, PhD)*, age, gender etc.
getting a particular illness	other illnesses, smoking, alcohol consumption, eating habits, lifestyle, hereditary disease risks, BMI, age, gender etc.
need for replacing hip or knee prostheses	illnesses, lifestyle, daily body movements, targeted body exercises, BMI, age, gender etc.
subjective well-being (individual happiness index)	financial situation *(heavy debt, moderate debt, no debt, properties)*, salary, total income, religiousness, unemployment, living in family *(yes, no)*, participation in public life, place of residence *(farmhouse, village, small town, big city)*, education *(no finished education, elementary school, secondary school, college, university, PhD)*, age, gender etc.

Applications areas (Goal functions)	**Predictors** (Characteristics of the individuals that influence these occurrences)
infection or death of experimental animals in medical and pharmaceutical experiments**	antecedent treatments/medications/surgeries, age
yield of a particular crop in experimental soil parcels**	soil chemistry, daily sunny hours, fertility, use of organic/*synthetic* fertilizers, soil moisture, irrigation rate, tillage etc.

Notes:

* This example is presented in more detail as the fifth case study in the last article of this special issue, titled *Illustrating real-life ATOM application case studies*.

** In these examples – unlike the earlier ones – the "individuals" are not even humans, but animals or soil parcels

A short prospective exposition of the content of the present special issue

Lajos Izsó: *The concept of an AI-based expert system (ATOM) for predicting job success* (this very article, at the same time also an issue editor's introduction)

Bence Gergely & Szabolcs Takács: *ATOM – a flexible multi-method machine learning framework for predicting job success*
ATOM's outstanding flexibility is primarily ensured by using expediently selected concurrent algorithms. This means that in ATOM, as opposed to the general practice of specifying a single model, several ML algorithms run in parallel. Thus, we can choose the solution that best suits the given situation. The main advantage of competitive algorithms is that they can adapt to the diversity of workforce selection. It is also adaptable to student datasets of variable size and quality, to expert evaluation, as well as to specific job characteristics and latent data generation processes. However, by increasing the flexibility of the procedure, we also increase the possibility of the so-called "overfitting", which also means that the algorithm only learns the data itself, i.e., it will not – or only to a limited extent – be able to generalize and identify patterns. We solved this problem in ATOM with proper cross-validation.

Judit T. Kárász & Szabolcs Takács: *Use of open and closed items in automation of evaluation systems*
Can we leave out open-ended items during automatic processing without significant distortion? Answering this question was decisive for the development of ATOM. Involving more than 80,000 respondents, we tested the mass consequences of omitting open-ended items on the data of the National Assessment of Basic Competencies (NABC). Our test runs showed that during the continuous evaluations, we were able to show quite close correlation levels – over 0.95 – between the versions that included open-ended questions and those that did not.

Examination of the results at the category level revealed that typically there are relatively significant differences at the two ends of the measurement scale, which has consequences for the application of ATOM as well.

Máté Pusker, Bence Gergely & Szabolcs Takács: *ATOM's structure – employee and employer feedback, survey site*
As described earlier, the ATOM program package has the following three main sets of functions according to the three main user categories that typically occur: (1) employees' functionalities, (2) employers' functionalities, and (3) experts' functionalities. The specific particular functions of these sets of functions can be accessed from the ATOM opening screen, via the following four primary windows: *"Users"*, *"Questionnaires"*, *"Setup"*, and *"Campaigns"*. The article deals with a detailed description of these functionalities from a practical point of view.

Lajos Izsó, Blanka Berényi & Máté Pusker: *Jointly applying a work simulator and ATOM to prevent occupational accidents and MSD through workforce selection*
The primary goal of this article is to present a promising concept using a work simulator (like ErgoScope) and ATOM combined.

The essence of this approach is to predict candidates' propensity for MSD-type *(Musculoskeletal Disorder)* occupational diseases and for causing or suffering workplace accidents based on ErgoScope measurements as inputs to ATOM.

The purposeful combination of ErgoScope with ATOM can have a "synergistic" effect that reinforces each other's impacts, significantly reducing the likelihood of MSDs and workplace accidents. To put it simply, we propose to apply the appropriate outputs of ErgoScope as inputs to ATOM.

Lajos Izsó, Blanka Berényi & Szabolcs Takács: *Illustrating real-life ATOM application case studies*
In this article, five specific, real-life case studies are presented based on the results obtained with the help of experts' functionalities of ATOM. The employees' and employers' functionalities were not involved in these studies.

All the predictors and parameters of job success in these case studies were entered into ATOM as external files. For simplicity, reliability and uniformity reasons, parameters of job success were given on two-point (i.e., binary) scales, where 1 = "less likely to be successful in the job", 2 = "more likely to be successful in the job". For characterizing the overall categorization performance of ATOM, the ROC curves and the Precision-Recall curves were used. As opposed to assessing overall categorization performance based on all possible cutoff levels, it is also shown how particular local compromises – between sensitivity (recall), specificity, and precision – can be found, if necessary, by purposefully selecting cutoff levels.

In all the articles within this special thematic issue, there is double referencing. The internal references (citations), relating to this special issue, come first. Later, separately, the external references follow in the usual way.

Összefoglaló

Egy munkahelyi beválás előrejelzésére szolgáló MI-alapú szakértői rendszer (ATOM) koncepciója

Háttér és célkitűzések: Jelen cikk egyfelől az ebben a különszámban megjelenő – a munkahelyi beválás mesterséges intelligencia (MI) alkalmazásán alapuló előrejelzését ismertető – többi cikknek egy rövid, előretekintő ismertetését adja, másfelől pedig egy általunk kifejlesztett konkrét implementációt is bemutat vázlatosan.

Azon túlmenően, hogy a cikk szélesebb pespektívában ismerteti az MI alkalmazásának lehetőségeit és korlátait az emberierőforrás-menedzsment (EEM) különböző területein – mint a toborzás, munkaerő-kiválasztás, alkalmazás, képzés, teljesítmény nyomonkövető mérése és menedzselése, bérezés, munkaügyi viszonyok, munkahelyi biztonság- és egészségvédelem stb. –, a szöveg egyúttal ezen különszám szerkesztőjének előszava is.

Az EEM-területeket támogató jelenlegi MI-alkalmazások elsősorban a toborzásra korlátozódnak, a többi területeken a cégek még nagyrészt hagyományos módszereket használnak. Minthogy a gyakorlat azt mutatja, hogy a HR munkatársak rendszerint alulteljesítenek a munkaerő kiválasztása során, elsősorban erre a feladatra fejlesztettük ki az ATOM (Alkalmasság Tesztelési/Osztályozási Modul) nevű MI-alapú rendszerünket. Az ATOM a kiválasztáson túlmenően a toborzást is képes hatékonyan támogatni, és valamilyen mértékben még a munkahelyi biztonság- és egészségvédelmet is.

Az ATOM alapfeladata „megtanulni” az alkalmas prediktorok (a beválás előrejelzésére alkalmas változók) és a releváns beválási kritériumok közötti kapcsolatot az adott munkakörre.

Az ATOM kiemelkedő hatékonyságot és rugalmasságot biztosító újszerű vonása, hogy a magjában sok tanuló algoritmus fut konkurens módon, és az ezek által produkált eredmények közül a rendszer mindig a legjobbat fogadja el..

Kulcsszavak: mesterséges intelligencia (MI), gépi tanulás, emberierőforrás-menedzsment (EEM), munkahelyi beválás, ATOM

References of this Special Issue

Gergely, B., & Takács, Sz. (2023). ATOM – a flexible multi-method machine learning framework for predicting job success. *Alkalmazott Pszichológia*, *25*(3), 15–30.

Izsó, L., Berényi, B., & Pusker, M. (2023). Jointly applying a work simulator and ATOM to prevent occupational accidents and MSD through workforce selection. *Alkalmazott Pszichológia*, *25*(3), 73–91.

Izsó, L., Berényi, B., & Takács, Sz. (2023). *Illustrating real-life ATOM application case studies. Alkalmazott Pszichológia*, *25*(3), 93–114.

References

Bangerter, A., Roulin, N., & König, C. J. (2012). Personnel selection as a signaling game. *Journal of Applied Psychology, 97*(4), 719–738.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, *9*(1), 9–30.

Eubanks, B. (2022). *Artificial Intelligence for HR. Use AI to support and develop a successful workforce*. Second Edition. Kogan Page Limited.

Henle, C. A., Dineen, B. R., & Dulffy, M. K. (2019). Assessing intentional resume deception: Development and nomological network of a resume fraud measure. *Journal of Business and Psychology*, *34*(1), 87–106.

König, C. J., & Langer, M. (2022). Machine learning in personnel selection. In S. Strohmeier (Ed.), *Handbook of research on artificial intelligence in human resource management* (pp. 149–167). Edward Elgar.

Kordon, A. K. (2020). *Applying Data Science. How to Create Value with Artificial Intelligence*. Springer Nature Switzerland.

Tasdemir, F. (2015). A Study for Developing a Success Test: Examination of Validity and Classification Accuracy by ROC Analysis. *Procedia – Social and Behavioral Sciences*, *191*(2015), 110–114.

Wheeler, A. R. & Buckley, M. R. (2021). *HR without People? Industrial Evolution in the Age of Automation, AI, and Machine Learning*. Emerald Publishing Limited.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

ATOM – A FLEXIBLE MULTI-METHOD MACHINE LEARNING FRAMEWORK FOR PREDICTING OCCUPATIONAL SUCCESS

Bence Gergely
ELTE Doctoral School of Psychology
Károli University of the Reformed Church in Hungary
gergely.bence98@outlook.com

Szabolcs Takács
Károli University of the Reformed Church in Hungary
takacs.szabolcs.dr@gmail.com

Summary

Background and Aims: Presenting the statistical fundamentals of ATOM and its concurrent algorithms, with particular respect to demonstrate the flexibility of the decision-making module.

Methods: Simulating different classification problems using the Scikit Learn machine learning program package. During these simulations, the sample size, the number of variables, the number of groups, the proportion of incorrect classifications, and the distance between the groups were systematically changed.

Results: Based on 180 datasets, the Multilayer Perceptron performed the best in about 52% of the cases, and the Support Vector Classifier came in second place. It was found that every method proved to be better than any other in at least one case, which means that if we are dealing with a company or job where the given problem arises, these procedures provide a more accurate result. In addition, profound differences between different parameters of the same procedure were observed.

Discussion: Considering that the job selection aims to filter the best candidates, the accuracy of all procedures increases and, in general, it was shown that ATOM's algorithms indicate a performance much above the expected value of random categorization.

Keywords: recruitment automation, machine learning, psychological testing, multi-method approach

DOI: 10.17627/ALKPSZICH.2023.3.17

Introduction

Recruitment aims to provide the employer with appropriate human resources as cost-effectively as possible. Therefore, the selected employees must be able to perform the necessary tasks and have the cognitive and behavioural competencies required by the job (Hmoud & Varallyai, 2019). An effective selection process consists of several interrelated sub-processes but generally starts with defining the necessary tasks and abilities for the job, then continues with searching for and assessing the candidates, and ends with contracting the new employee (Ployhart, 2006). In the outlined process, human activity is essential since the selection cannot be, or is difficult to generalise. For the same reason, cognitive biases and heuristics decision-making are deeply rooted in selection (Whysall, 2018, Soleimani et al., 2022). For this reason, companies increasingly use recruitment software and attempt to partially or fully automate recruitment (Hmoud & Varallyai, 2019; Soleimani et al., 2022; Gonzalez et al., 2022; Liem et al., 2018).

Explanatory and predictive models

The primary goal of psychological science is to understand human behaviour (Yarkoni & Westfal, 2017). So, psychology primarily wants to explain phenomena with the simplest and most parsimonious models possible while placing less emphasis on prediction. So, in the vast majority of cases, psychology acts based on Occam's razor in model and theory formulation, i.e., it uses the most straightforward model with good explanatory power. The consequence is that the results can be only generalised within a closed theoretical framework and often have negligible predictive power (Robinaugh et al., 2021). In contrast, machine learning methods (especially deep neural networks) aim to maximise the prediction accuracy of the models. At the same time, mostly they do not provide an understandable explanation for how the phenomenon works (Yarkoni & Westfal, 2017). In that case, although it will provide a precise prediction for the given phenomenon, we will not (necessarily) know which variables and to what extent played a role in the outcome. Applied psychology often works with complex systems; therefore, the explanation of the processes is usually not the goal, mainly due to the scarcity of time and resources. Instead, the focus is on decision-making. Machine learning methods gained popularity in psychology, aiming to help professionals make decisions, such as in clinical diagnostics (Dwyer et al., 2018; Coutanche & Hallion, 2019). In the analysis of psychological experiments (Koul et al., 2018), academic success (Halde et al., 2016), and labour success (Liem et al., 2018).

The question is, do we want to understand the role of the factors involved during recruitment, or do we only want to provide a prediction? Suppose we only keep the explanation in mind. In that case, our selection process will probably be inflexible and not generalisable to other jobs, but we will earn a good understanding of the job's requirements. On the other hand, if we keep the prediction in mind and select the examined variables well in our model, our prediction will be accurate. However, if the variables are not appropriate, we will be unable to correct the prediction's inaccuracy.

In an ideal recruitment framework, one can optimise both aspects simultaneously:

giving a good prediction and indicating which variables play a role in the prediction come hand in hand (Kárász & Takács, this special issue).

Unique properties of recruitment data

The data arising from recruitment can be classified as ‘soft’ data. Its variability is much greater than data from physical measurement tools (Tannahil, 2007). Many times, this measurement error masks the otherwise complex data generation process. Due to the uncertainty of the variables, even complicated processes can appear linear (Yarkoni & Westfal, 2017). In order to reduce this uncertainty, work simulators and other instruments closer to physical measuring devices are often used in the field of work psychology (e.g., ErgoScope [Izsó, Berényi, & Takács, this special issue]).

The fact that it is difficult to access a large amount of data under given working conditions also contributes to the bias. Filling out long questionnaires can take a given employee out of production for several hours – however, it is difficult to make good predictions from a small amount of data (Yarkoni & Westfal, 2017). If going through the test battery takes a long time, missing data and systematic distortion of the test result occur more often (Nagybányai Nagy, 2013). That is why it is crucial to only ask for data that is needed – but we can only determine optimal test battery from preliminary measurements (Kárász & Takács, this special issue).

In addition, the quality of the data can also be questionable. There are often no established criteria for evaluating the performance of employees (Maji and Bera, 2020; van Esch et al., 2019). In many jobs, it is impossible to use objective performance indicators, and we can only obtain performance measures based on the subjective evaluation of HR professionals (Kárász & Takács, this special issue). Often, the selected psychological scales do not have predictive power, even in the case of high-reliability performance evaluations. The use of measurement tools is often limited to the kind of psychological tests the job has access to and whether they evaluate the effectiveness of the tests in the given recruitment process (Izsó, Berényi, & Takács, this special issue). The strength of the predictions largely depends on the quality of the input data. Analyses with low-quality data can raise serious validity problems – but these can also be handled to a significant extent by using different, more robust statistical procedures (Gergely & Vargha, 2021). It is often difficult to determine the quality of the data, but the multi-method approach adopted during the replication crisis can help a lot in drawing accurate conclusions. The essence of the multi-method methodology is that a given number of adequate statistical procedures are performed on a statistical question, then the obtained results are aggregated, thus making a more robust decision.

At the same time, the disadvantage of systems using more robust or complex methods is that the output data are difficult to interpret by professionals. Interpretability can be helped by providing the minimum information necessary for decision-making. For example (Izsó, Berényi, & Takács, this special issue) found that the feedback on the order of applicants and their classification into only two discrete acceptance categories can be sufficient for making a decision.

Decision-making module of ATOM framework

ATOM is a modular web-based framework that includes the compilation of the test battery, the organisation of recruitment campaigns, the analysis of the results, and the provision of automatic psychological feedback (Kárász & Takács, this special issue). Due to this structure, the goal of ATOM is to reduce the need for human resources in recruitment campaigns, thereby becoming a cheaper and more convenient alternative to traditional testing.

Training and test data requirements

ATOM's decision-making module is a flexible machine-learning framework combining several statistical methods. In each case, the input data consist of subscales of psychological and performance tests that have been validated and have high reliability. The subscale score is given by the sum of the items weighted to the subscale, which is standardised before the analyses (with a mean of 0 and a standard deviation of 1). The purpose of standardisation is to make the different subscales comparable, which is often a prerequisite for applied machine learning algorithms (Kárász & Takács, this special issue).

We need two types of input data to use the decision-making module: a training and a testing data file. In this case, the testing dataset represents the questionnaire results of the individuals applying for the given position, while the training dataset can be obtained from two sources. In the training data, we need information about whether an individual was proven to be a suitable candidate for a given job.

If the recruiting company has many employees, we can obtain the training data from the questionnaires filled out by these employees. Then these results must be labelled. Labelling means that the employees participating in the testing are classified into one of the predefined discrete groups (i.e., suitable, conditionally suitable, or not suitable for the position). These discrete groups can be created based on more objective performance measures (e.g., how many partners a sales employee contracts within a year), or the subjective evaluation of specialists can also provide the labelling. Expert evaluation is often fraught with cognitive biases, so to create optimal labelling, we need to ask for the opinions of several independent experts (Hallgren, 2012). Of course, the phenomenon of 'garbage in, garbage out' arises here, i.e., if the algorithms are trained with low-quality data, then the result (classification) will also be of poor quality. It is important to note that the decision-making module is structured in such a way that we can indicate the quality of the classification and the importance of the psychological and performance variables used, thus improving the efficiency of labelling and testing in the future.

If the recruiting company has few employees or there is no time for testing and labelling employees, then the quantification of expert opinions is a possible direction. Hmoud and Varallyai (2019) emphasise that the first step of the recruitment process is analysing the given job and assessing the necessary competencies. Hence, HR experts and work psychologists have a professional profile and optimally choose measurement instruments for this professional profile. ATOM's decision module can quantify this professional profile based on the measurement tools. First, the experts indicate which variables are important, moderately important, or not important for the given job and also define the results required to be classified in the suitable candidate category. After that, we create a mixture of multidimensional normal distributions, which is parameterised based on the given expert values, whereas the non-determinable parameters (e.g., covariance) are fixed based on several different models (Gergely & Vargha, 2021). Labelling is defined here by belonging to a given component of the mixture distribution. The resulting artificial datasets reflect the expert opinion, but at the same time, they also include the uncertainty of the expert opinion.

It is important to note that the two student data file types cannot only operate independently of each other. For example, it may be the case that the company has few employees, but we take the tests with them, but to have a sufficient number of items, we also take the expert opinion into account.

Concurrent algorithms, hyperparameters, and cross-validation

If we have surveyed the employees or created the learning datasets, the next step is to fit the selected algorithms to the data. During the data analysis phase, the algorithms must predict the labels defined in the learning dataset, and the quality of the algorithm is determined by the accuracy of this prediction. The main idea behind ATOM's decision-making module is the use of concurrent algorithms, i.e., in contrast to the general practice (which specifies a model for the given use), several machine learning algorithms run in parallel, and the goal is to select the best solution for the given situation. The main advantage of competing algorithms is that they can adapt to the diversity of workplace selection, training data of varying size and quality, expert evaluation, and the specific characteristics of the job and latent data generation processes.

In order to optimally use and evaluate multiple algorithms together, three steps are required: hyperparameter setting, cross-validation, and measurement of the prediction accuracy.

Hyperparameters are the values that influence how a given algorithm works. Different algorithms can have different hyperparameters, and it is usually impossible to determine a combination of values that gives the best result in every case. In order to make it possible to measure which setting is the most optimal, we defined a hyperparameter space for each algorithm, with which we can determine which hyperparameter setting is the most suitable for the given problem by testing the algorithm with all possible hyperparameter combinations.

Some of the algorithms are not flexible. Logistic regression, being a generalised linear model, can fit one kind of function (a sigmoid function), while neural networks with different parameterisations can use many different non-linear functions. To take advantage of the strengths of the different algorithms, we use the method used for hyperparameter setting in this case as well. We create a model and hyperparameter list, the combination of which we fit the data and measure their effectiveness.

Machine learning algorithms learn based on how accurately they can predict the training dataset's labels. By increasing the flexibility of the procedure, we increase the possibility of overfitting. Overfitting means that the algorithm only learns the data, i.e., it will not be able to reveal general patterns so that it will provide a suboptimal prediction in the case of previously unseen data. To minimise this possibility, we performed cross-validation on the entire model and hyperparameter space. The essence of cross-validation is to randomly divide the learning dataset into n equal parts and then create all possible (i.e., n pieces) partial learning datasets. The partial learning sets consist of n – 1 equal part, and the quality of the algorithm is tested only on the remaining one data part. This way, we test the algorithm's effectiveness on data that it has never seen before. We perform this process on all (n pieces) of the learning data set and then average the accuracy of the prediction, thus obtaining an estimate of how well the given algorithm performs on data it has not yet seen.

So far, we have not precisely defined what we mean by the efficiency of the algorithm and the quality of the prediction. There are several measures for this, depending on what we want to maximise/minimise in the given application. In this study, for the sake of simplicity, we used the percentage of correctly classified cases as an efficiency indicator. The percentage of correctly classified cases measures the percentage of predicted labels that match the actual labelling. In real selection situations, it makes sense to use several efficiency indicators, as the goal is usually not to categorise all applicants accurately but to filter out the best applicants. They will be forwarded to the interview process. In this case, a good efficiency indicator can be the percentage of correctly classified cases or the rate of false positives in the suitable candidate category. In summary, during the analysis phase, we select the algorithm-hyperparameter combination that receives the best score in the cross-validation procedure based on our determined efficiency measure.

Selected algorithms

During the construction of the decision-making module, the Python programming language was used in combination with the open-source Scikit-Learn program package (Python, 2021; Pedregosa et al., 2011).

Since we defined our dependent variable as a discrete category, we chose supervised learning algorithms that can solve classification problems. The complexity of the algorithms and the fact that the selected algorithms use different heuristics also played a role in the selection.

ATOM's decision-making module currently supports Logistic Regression (Wright, 1995), its regularised version (Cherkassky & Ma, 2003), the Support Vector Classifier algorithm family, Random Forest (Breiman, 2001), Adaboost (Freund & Schapire, 1997) and Multilayer Perceptron (Collobert et al., 2004).

Both the advantage and disadvantage of Logistic Regression lie in its simplicity: it is a generalised linear model capable of solving classification problems and requires few parameters for its operation. The Support Vector Classifier is a family of algorithms effective for multi-dimensional problems, even when the number of variables is larger than the number of sample elements. In addition, it is flexible since the function used for decision-making can be influenced by using different kernels. The disadvantage is that the probability of overfitting increases in the case of many variables. In such cases, regularisation and cross-validation should be used. Random Forest and Adaboost are ensemble methods that combine several simple prediction algorithms (typically decision trees). While Random Forest is an averaging method that builds decision trees independently and then aggregates their results, Adaboost makes sequential estimates, i.e., builds a more efficient one from several weaker classification algorithms. Finally, the Multilayer Perceptron belongs to the family of artificial neural networks (ANN), but its version used in ATOM has only one hidden layer. The advantage of this solution is flexibility, its disadvantage is that it needs to estimate the weight and bias of the edges, which depends on the width of the input, output, and the hidden layer.

Output data and model evaluation

The final step in the decision-making module is to provide the output data. The most basic output data is the predicted labelling, and how the algorithms categorised the applicants. In some cases, this may be sufficient to select the applicants who enter the interview process, but the disadvantage is that it does not indicate how uncertain the decision was. The uncertainty of the decision can be quantified with labelling probabilities. When calculating the labelling probability, we do not classify the applicants under a label but give the probability of belonging to each group. For example, let us take two applicants; both of them were classified in the suitable category, but when we examine the labelling probability, we see that one belongs to the successful group with a 90% probability, while the other only with 65%.

In addition, we need to use measures that provide information about the performance of the models. Since not all methods can test the significance of the variables or indicate their importance, we used the Shap-value method (Shapley & Snow, 1952; Bowen & Ungar, 2020), which estimates the contribution of each variable to the prediction.

The purpose of this study is to demonstrate the flexibility of ATOM's decision-making module using simulations. In the case of different types of data occurring in the selection, the advantage of using several methods together prevails. So, in the case of simulated datasets, there will be at least one time when the given algorithm family gives the most accurate estimate, and the accuracy of the estimates will be similar between the models.

Methods

After the literature presentation and ATOM's methodology, the question may arise: Why is it necessary to use several concurrent process algorithms? In the machine learning literature, researchers traditionally present one procedure and compare it with algorithms created for a similar purpose or application. In this research, we want to show that using several simpler procedures (with fewer parameters) can achieve the robustness necessary to use data from psychological testing for recruitment.

The system is analysed using a simulation study. We created different classification problems during the simulation using the Scikit Learn machine learning program package (Pedregosa et al., 2011; Python, 2021). When creating the classification problems, we changed the size of the sample, the number of variables, the number of groups, the proportion of incorrect classifications, and the distance between the groups.

The sample size was 50, 100, 200, 500 and 1000, respectively, meaning that the total sample size for the training dataset was one of the values above. We considered that the sample size in psychology is often small and rarely exceeds 1,000 people. In addition, the sample size also reflects the number of individuals who can be tested on the Hungarian labour market; usually, medium-sized enterprises have around 50 employees, and in the case of large companies, it is not uncommon for a workforce of over 1,000 people (KSH, 2018).

The number of variables was divided into two categories: explanatory and redundant. Explanatory variables are those that can significantly predict which group the test person belongs to, while redundant variables are those that have no predictive power. The number of generated explanatory variables was 5, 10 and 20, respectively, for which we also created 10 redundant variables in each case. Redundant variables were considered important because it is common in workplace selection that some performance indicators do not have direct predictive power for the given job and are often used only because they are available or included in the test battery used by the company. An important question, in this case, is whether our automated procedures can filter these redundancies, thereby providing information about which variables should be used in the future.

The number of groups, i.e., the defined classes, was 2, 3 and 4. Here, we found that in practice, the inaccuracy of the grading increases as the number of categories increases in most companies. This is because the 2-point scale usually carries the essential information (suitable, not suitable candidate), and the 3-point scale (suitable, conditionally suitable, not suitable).

Table 1. Different parameters of the simulation setups

Parameters	**Values**				
Sample size	50	100	200	500	1000
No. variables	5	10	20		
Redundant variables	10				
No. groups	2	3	4		
Proportion of incorrect classifications	0.01	0.1			
Distance between groups	1	0.75			
Total	**180 classification problems**				

Source: created by the authors based on simulation details

We used the so-called incorrect classification ratio (0.01, 0.1), which means that 1% and 10% of the cases are already included incorrectly in the training dataset. Partly due to the inaccuracy of the suitability scale mentioned in the previous paragraph and partly due to the heuristic nature of human classification, we used these incorrect classification rates since it is assumed there are also false groupings in real datasets. This allows us to test the extent to which the learning algorithms can correct these evaluation biases.

The distance of the groups was set to 1 and 0.75, which means how ‘separated’ the clusters are from each other. A larger value means more separation, which results in an easier classification problem, while a smaller value means less separation and a more difficult classification problem. In a system where there are significant differences between suitable and not suitable candidates (1 standard deviation), we can expect significantly better results than in a case where the difference between them is smaller (0.75 standard deviations). Here, this should be understood as the number of standard deviation differences between the mean values when creating the mixture distributions.

We simulated a classification dataset with all possible combinations of these parameters, resulting in 180 different problems. Then, we ran the algorithms of ATOM framework with different parameterisations on each data file.

The effectiveness of the different algorithms and their different parameterisations was measured by the average accuracy of the classification (number of correct classifications / all cases). In this study, we used the first version of the ATOM, which only included accuracy as an outcome measure. For each algorithm, we present the number of cases when the given method provided the best accuracy. Moreover, we report the rank means over all 180 simulations.

To test and visualise the performance differences between algorithms and their different parameterisations, we perform a Kruskal-Wallis test with Bonferroni corrected pairwise comparisons and present the accuracy's median and the median absolute deviance.

Results

First, we consider the runtime of the simulations. The total runtime of the simulations study was approximately 11 minutes.[1] For the smaller datasets (50, 100) the grid search algorithm took only a few seconds (1-5s), none of the larger datasets took longer than 30 seconds to finish. Support Vector Classifier was the slowest to fit, albeit having the most parameters to sweep through with the grid search algorithm. Overall, we think that the speed of the algorithm is more than adequate for its use cases.

In the first step, we present which algorithm provided the most accurate prediction across all 180 datasets. In 51.1% of the cases, the Multilayer Perceptron, i.e., the neural network with one hidden layer, provided the best prediction, and the Support Vector Classifier came in second place. It is important to note here that the different parameterisations were not considered, the ratios here show how many times the given procedure provided the best prediction regardless of the different settings.

[1] All the simulations were running on a 2022 Apple Macbook Pro with an M1 Pro chip. The used packages were available for arm-type systems.

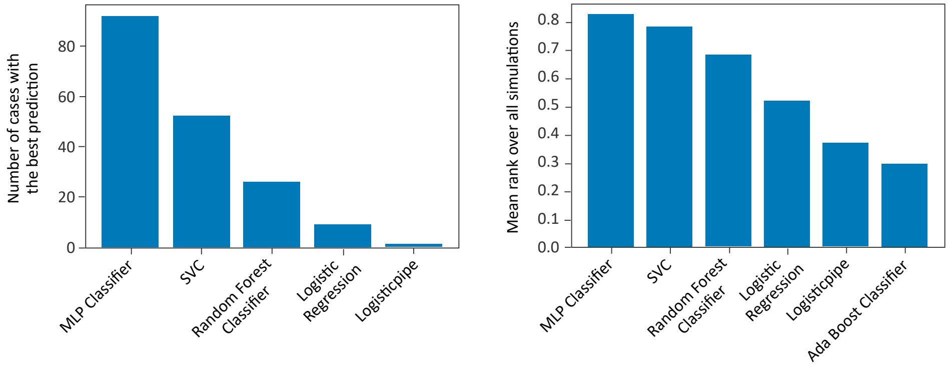


Figure 1. Number of cases when each algorithm provided the best prediction and the mean rank of the algorithms over all simulation

Source: the results were calculated and visualised using Python

However, the simulation aimed to show cases where it is unclear which procedure to choose. There were 10 classification problems each where Logistic Regression and the Logistic Pipe gave the most accurate prediction.

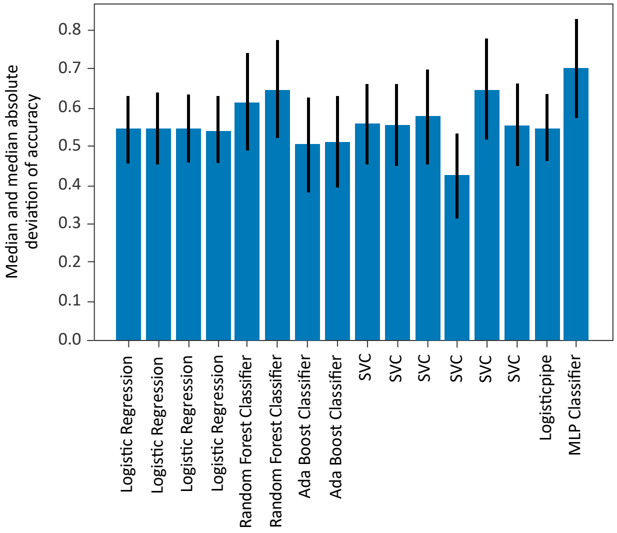


Figure 2. Median and median absolute deviation of accuracy for all algorithms and their parameterisations over all simulations

Source: the results were calculated and visualised using Python

That is, all methods except Adaboost proved to be better than any other in at least one case. This means that if we deal with a company or a job where the given problem arises, these procedures provide a more accurate result. However, the actual data cannot be analysed based on the simulation aspects since we usually have no information which procedure will be the most suitable before the analysis.

At the same time, we also examined the median accuracy of the procedures on all datasets and their median absolute deviations. There can be big differences even between different parameters of the same procedure. Based on the Kruskall-Wallis test [H(5) = 111,656; p < 0.001], and the pairwise comparisons, there are significant differences in the performance of the algorithms[2], and ultimately the algorithms can be ranked as Multilayer Perceptron, SVC, Random Forrest, Logistic Regression, Logistic Pipe and Adaboost, respectively.

Most importantly, all average results are typically above the 0.5 bands. This means that even in the case of a 2-valued prediction (success/failure), the prediction procedures have better results than completely arbitrary decision-making.

However, it is rarely important to classify each applicant accurately on the employer's side. It is much more important how well the given algorithm can guess the top 10% of applicants (the best 5–10–20 applicants), as these candidates will typically be the ones who will participate in the interview process.

Thus, we also looked at the median and median absolute deviations of the percentage of correctly classified cases for the top 10% of employees. In this case, the best method was the neural network: with a mean percentage of correctly classified cases of 70% and a standard deviation of 28%. So, if we are only interested in who the experts are, we can show an acceptable accuracy (in the case of 2 categories, we can show a rate of well over 50%).

Table 2. Post hoc comparison with Bonferroni correction

Post hoc comparisons - Accuracy							
		Mean Difference	SE	t	Cohen's d	p_{tukey}	p_{bonf}
AdaBoost-Classifier	Logistic-Regression	-0.046	0.011	-4.245	-0.274	< .001	< .001
	Logisticpipe	-0.047	0.015	-3.095	-0.283	0.024	0.030
	MLPClassifier	-0.141	0.015	-9.255	-0.845	< .001	< .001
	RandomForest-Classifier	-0.088	0.012	-7.074	-0.527	< .001	< .001
	SVC	-0.035	0.010	-3.477	-0.212	0.007	0.008

[2] Based on the Shapiro-Wilk test, the accuracies were likely non-normal in all cases.

Post hoc comparisons - Accuracy							
		Mean Difference	SE	t	Cohen's d	p_{tukey}	p_{bonf}
Logistic-Regression	Logistic-pipe	-0.001	0.014	-0.103	-0.009	1.000	1.000
	MLPClassifier	-0.096	0.014	-6.850	-0.571	< .001	< .001
	RandomForest-Classifier	-0.042	0.011	-3.924	-0.253	0.001	0.001
	SVC	0.010	0.008	1.297	0.062	0.787	1.000
Logistic-pipe	MLPClassifier	-0.094	0.018	-5.334	-0.562	< .001	< .001
	RandomForest-Classifier	-0.041	0.015	-2.681	-0.245	0.079	0.111
	SVC	0.012	0.013	0.881	0.071	0.951	1.000
MLP-Classifier	RandomForest-Classifier	0.053	0.015	3.479	0.318	0.007	0.008
	SVC	0.106	0.013	7.865	0.633	< .001	< .001
RandomForest-Classifier	SVC	0.053	0.010	5.187	0.316	< .001	< .001

Note: P-value adjusted for comparing a family of 6

Source: the results were calculated using JASP (Love et al., 2019)

Discussion

In this study, we presented the decision-making module of the ATOM framework and the advantage of the competitive algorithms method with the help of a simulation study.

ATOM's decision-making module was designed to answer the questions outlined in the introduction, namely the uncertainty coming from psychological assessment in recruitment scenarios. This uncertainty is often due to the small amount of data available for a given position. In case of a small sample size ATOM can quantify the expert evaluation of the different suitability categories and create a mixed dataset of actual and simulated candidates. Since companies rarely provide objective labelling of their employers, ATOM supports the good practice that HR experts independently assess the suitability of the employers. In the expert evaluation process, it is worth expecting high interrater reliability before starting the analysis. The goal is therefore not to exclude HR professionals from recruitment – but to best allocate their capacity and make the most optimal use of their expertise in the pre-screening and interview phase. In the case of longer-term cooperation, the amount of recruitment data is increasing over time, thus the prediction will be gradually better, but the quantification of expert opinion can reduce the cold start period, where our predictions are less than adequate. If a company is using ATOM for a longer period, it will gather data about not only the suitable candidates, but also the candidates who have not met the expectations. This way, the training sample becomes more representative of the suitability categories; therefore, the algorithms will be more accurate overall.

It is important to note that the goal is not to accurately predict all categories of candidates. The goal is, instead, to identify who are the most suitable and likely to succeed, from whom the company can select the best candidates for the interview process. This process is facilitated by ATOM's decision-making module by freely changing the efficiency measures, thereby tailoring the analysis to the expectations of the job.

We selected algorithms that are not based on the same mathematical background; they require different assumptions and have varying robustness. That is, while the Support Vector Classifier is sensitive to the kernel type, it achieves good results in cases where the number of variables used is high. AdaBoost is not sensitive but tends to overfit in the case of many variables. At the same time, the power of the decision-making module is manifested in the fact that we do not have to take these assumptions into account, since these algorithms compete with each other on the training dataset, with different parameterisations and automatic model selection.

To account for the uncertainty of the outcomes, instead of just presenting the predicted suitability, we also report the probabilities of belonging to each category. In this way, employers can create their own rankings: filtering the least likely succeeding candidates or selecting the most potent ones (Izsó, Berényi, & Takács, this special issue).

Based on the simulation, we can say, that in the case of our developed system, the selected algorithms create a flexible framework. Moreover, all algorithms, except the Adaboost provided the best prediction in at least one case. Nevertheless, it was expected that the neural network would produce the best results due to the algorithm's robustness (Collobert & Bengio, 2004).

Concerning the accuracy of the prediction, we did not experience any substantial differences between the different methods and their different parameterisations. The average performance was above 50% respectively. Note that the expected value of a completely random selection is 35%. So, each algorithm results in a much more accurate categorisation on average. If we consider that the job selection aims to filter the best candidates, the accuracy of all procedures increases and, in general, our algorithms show a performance well above the expected value of random categorisation (rate of 35%).

The current algorithm's limitation is that it selects a single model in each case and does not account for the strength of different models. A model selection resembling the Bayesian model averaging would be more suitable than choosing the most accurate model. Furthermore, the algorithm's flexibility needs to be further assessed with different types of data and jobs.

The limitation of the simulation study is that the simulated datasets all came from a mixture of normal distributions, with equal distances between the centroids of the components.

Összefoglaló

ATOM – egy rugalmas, több módszert alkalmazó gépi tanulási keretrendszer a munkahelyi bevális előrejelzésére

Háttér és célkitűzések: Jelen kutatás bemutatja az ATOM szoftvert és annak statisztikai megfontolásait, különös tekintettel a döntéshozatali modul rugalmasságának demonstrálására. *Módszer*: Scikit Learn segítségével különböző osztályozási problémákat szimuláltunk. A szimulációk során szisztematikusan változtattuk a minta méretét, a változók számát, a csoportok számát, a hibás osztályozások arányát és a csoportok közötti távolságot. *Eredmények*: A 180 szimulált adatállomány alapján a Multilayer Perceptron az esetek mintegy 52%-ában a legjobban teljesített, a második helyen pedig a Support Vector Classifier végzett. Megállapítottuk, hogy minden módszer legalább egy esetben jobbnak bizonyult a többinél, ami azt jelenti, hogy ha olyan céggel vagy munkakörrel foglalkozunk, ahol az adott probléma felmerül, akkor ezek az eljárások pontosabb eredményt adnak. Ezenkívül lényeges különbségeket figyeltünk meg ugyanazon eljárás különböző paraméterezései között. *Következtetések*: Tekintettel arra, hogy a kiválasztás célja a legjobb jelöltek kiszűrése, az összes eljárás pontossága növekszik, ha csak a legegyertelműbben kategorizálhatókat keressük. Általánosságban megmutatkozott, hogy az ATOM algoritmusai a véletlenszerű kategorizálás várható értékét jóval meghaladó teljesítményt jeleznek.

Kulcsszavak: munkaerő-kiválasztás automatizációja, gépi tanulás, pszichológiai tesztelés, konkurens algoritmusok alkalmazása

References of this Special Issue

Izsó, l., Berényi, B., & Takács, Sz. (2023). Illustrating real-life ATOM application case studies. *Alkalmazott Pszichológia*, *25*(3), 93–114.

References

Bowen, D., & Ungar, L. (2020). Generalized SHAP: Generating multiple types of explanations in machine learning. *arXiv*. https://arxiv.org/abs/2006.07155

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cherkassky, V., & Ma, Y. (2003). Comparison of model selection for regression. *Neural Computation*, *15*(7), 1691–1714. https://doi.org/10.1162/089976603321891864

Collobert, R., & Bengio, S. (2004). Links between perceptrons, MLPs and SVMs. *Proceedings of the Twenty-first International Conference on Machine Learning*, 23. https://doi.org/10.1145/1015330.1015415

Coutanche, M. N., & Hallion, L. S. (2020). Machine learning for clinical psychology and clinical neuroscience. In A. G. C. Wright & M. N. Hallquist (Eds.), *The Cambridge Handbook of Research Methods in Clinical Psychology.* Cambridge University Press. https://doi.org/10.1017/9781316995808.041

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Gergely, B., & Vargha, A. (2021). How to Use Model-Based Cluster Analysis Efficiently in Person-Oriented Research. *Journal for Person-Oriented Research*, *7*(1), 22–35. https://doi.org/10.17505/jpor.2021.23449

Gonzalez, M. F., Liu, W., Shirase, L., Tomczak, D. L., Lobbe, C. E., Justenhoven, R., & Martin, N. R. (2022). Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes. *Computers in Human Behavior, 130*(May), 107179. https://doi.org/10.1016/j.chb.2022.107179

Halde, R. R., Deshpande, A., & Mahajan, A. (2016). Psychology assisted prediction of academic performance using machine learning. *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 431–435. https://doi.org/10.1109/RTEICT.2016.7807857

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–43. https://doi.org/10.20982/tqmp.08.1.p023

Hmoud, B. I., & Várallyai, L. (2020). Artificial intelligence in human resources information systems: Investigating its trust and adoption determinants. *International Journal of Engineering and Management Sciences*, *5*(1), 749–765. https://doi.org/10.21791/IJEMS.2020.1.65

KSH (Központi Statisztikai Hivatal) (2018). Munkaerőpiaci helyzetkép, 2014–2018. http://www.ksh.hu/docs/hun/xftp/idoszaki/munkerohelyz/munkerohelyz17.pdf

Koul, A., Becchio, C., & Cavallo, A. (2018). PredPsych: A toolbox for predictive machine learning-based approach in experimental psychology research. *Behavior Research Methods*, *50*(4), 1657–1672. https://doi.org/10.3758/s13428-017-0987-2

Liem, C., Langer, M., Demetriou, A., Hiemstra, A. M., Sukma Wicaksana, A., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 197–253). Springer. https://doi.org/10.1007/978-3-319-98131-4_9

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., ... & Wagenmakers, E. J. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*, 1–17. https://doi.org/10.18637/jss.v088.i02

Nagybányai Nagy, O. (2013). *The Effect of Response Style Characteristics on the Measuring Efficiency of Self-administered Testing Methods* [Doctoral dissertation]. Eötvös Loránd University.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Ployhart, R. E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management*, *32*(6), 868-897.

Python, W. (2021). Python. *Python Releases for Windows*, *24*.

Robinaugh, D. J., Haslbeck, J. M., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743. https://doi.org/10.1177/1745691620974697

Shapley, L. S., & Snow, R. N. (1952). Basic solutions of discrete games. *Contributions to the Theory of Games*, *1*, 27–35. https://doi.org/10.1515/9781400881727-004

Soleimani, M., Intezari, A., & Pauleen, D. J. (2022). Mitigating cognitive biases in developing AI-assisted recruitment systems: A knowledge-sharing approach. *International Journal of Knowledge Management*, *18*(1), 1–18. https://doi.org/10.4018/IJKM.290022

Tannahill, G. K. (2007). *A study of soft skills for IT workers in recruitment advertising.* Capella University

van Esch, P., Black, J. S., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, 215–222. https://doi.org/10.1016/j.chb.2018.09.009

Whysall, Z. (2018). Cognitive biases in recruitment, selection, and promotion: The risk of subconscious discrimination. In V. Camen, & S. Nachmias (Eds.), *Hidden inequalities in the workplace: A guide to the current challenges, issues, and business solutions* (pp. 215–243). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-59686-0_9

Wright, R. (1995). Logistic regression. *Reading and Understanding Multivariate Statistics*, 217–244.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

USE OF OPEN AND CLOSED ITEMS IN AUTOMATION OF EVALUATION SYSTEMS

Judit T. Kárász
ELTE Eötvös Loránd University, Doctoral School of Education
ELTE Eötvös Loránd University, Institute of Education
Károli Gáspár University of the Reformed Church in Hungary
t.karasz.judit@ppk.elte.hu

Szabolcs Takács
Károli Gáspár University of the Reformed Church in Hungary
takacs.szabolcs.dr@gmail.com

Summary

Background and Aims: The design of automated evaluation systems raises the problem of open-ended questions or tasks that require living labour. Coding open-ended questions is a costly, time- and labour-intensive task. Reviewing and selecting CVs reduces the amount of time spent on face-to-face interviews. Our research question is: which subjects are affected by the omission of open-ended questions, and what are the consequences for evaluating results? Our study was conducted on data from the National Assessment of Basic Competencies, which can also be understood as an assessment system currently in the automation process.
Methods: The original proportions were restored by weighting according to the measurement methodology. In this study, we compared achievement scores and proficiency levels calculated based on the whole test booklet and the basis of closed items only.
Results: Ability scores calculated from the entire test and closed items show a strong correlation.
Discussion: Our calculations demonstrate that open-ended items are needed in ability ranges where fewer items are available in the first place. By omitting the open-ended items, a significant "loss" is typically incurred by those for whom we have less information, who are classified as "very high performers" or "very low performers".
Keywords: automated evaluation, workplace validation, evaluation system, test format, National Assessment of Basic Competencies

DOI: 10.17627/ALKPSZICH.2023.3.33

Introduction

Previous studies presented in this thematic issue (Gergely & Takács, this special issue) have shown the potential benefits of automated evaluation systems. In our view, this does not mean that professionals are not needed. On the contrary, the time freed by automated systems can be used by professionals to perform tasks requiring greater expertise in a more focused way.

In the field of education, there is debate regarding the difference between closed and open-ended questions that require post-coding, for example, during international student performance measurements (e.g., Bingölbali & Bingölbali, 2021; Lafontaine & Monseur, 2009). How can the assessment of the latter be automated (Çınar et al., 2020; Yamamoto et al., 2017), or how much can we rely solely on the information in the closed questions? Although large-scale student assessments show us a distant analogy with workplace selection by content, we can see a considerable analogy at the level of mathematical structure. In both situations, proficiency levels are defined along predetermined ability scales, for which levels well-characterized abilities and expected performances can be formulated (Balázsi et al., 2014; OECD, 2019). Thus, the decision-making, that each test subject is classified into levels based on the measured ability, and some kind of expected performance can be associated with the ability, can be interpreted, and treated in an analogous way to workplace selection. In addition, the Organization for Economic Co-operation and Development (OECD) and the Program for International Student Assessment (PISA) examines the 15-year-old population with its literacy-based test in school conditions, because the tasks of the assessment measure "the existence of knowledge and skills that are essential for full participation in modern societies" (OECD, 2019, p. 13). The National Assessment of Basic Competencies (NABC) assesses all 6th, 8th, and 10th-grade students in Hungary and follows the OECD PISA assessment in the main lines of content and methodology (Auxné Bánfi et al., 2014).

In the case of workplace surveys, we rarely experience such dimensions and sample sizes as in the case of large-scale student assessments. In the case studies section of the thematic issue (Izsó, Berényi & Takács, this special issue), samples of a maximum of a few hundred people can be found, and in the case of the NRSZH data, the sample size was 15,000 people, which is also considered extreme. The National Assessment of Basic Competencies provides a great volume of participants and information, moreover, it is a regular assessment and not a one-time measurement. In this sense, testing on the data set of the National Assessment of Basic Competencies can be said to be extreme compared to the number of workplace assessment tests.

Let us take an example of a more extreme measurement, but one that occurs every year. The National Assessment of Basic Competencies (Belinszki et al., 2020) is conducted in Hungary every year in 3 grades (6th, 8th, and 10th), with about 80,000 students per grade. Students complete a test consisting of two test sections, each of which consists of approximately 50-60 questions. Alarge proportion of the questions are simple or multiple-choice, while a smaller proportion, in the order of one-third (Balkányi et al., 2018; Lak et al., 2018), are open-ended, i.e., they require the students to construct the answer independently. An open-ended question may be one that is an open question (how many apples have been picked and a number is expected). However, in the case of computer-assisted data collection, a computer can assess it quickly. An open-ended question should be coded if "live" processing is essential, including mathematical reasoning, proof, or a composition response in a reading comprehension test (Balázsi et al., 2014). In the case of paper-and-pencil tests, both forms of open-ended questions should be coded.

We suppose that during a computerised data recording, only about 10% of the 100–120 questions are open-ended – or even coded. To make the calculation easier, let us assume that there are 50 questions in 2 fields of knowledge, 10% of which are to be coded, so a total of 10 questions needs to be read through. These fields are divided into 3 or 4 content areas each, measured with both closed and open-ended questions. The closed questions are coded and computer-evaluated after scanning, while the answers to the open-ended questions are evaluated and coded by experts after multiple rounds of training. The ability scores are calculated separately into mathematics and reading comprehension scores, so content areas are combined into an aggregated indicator.

Ten questions are not that many compared to 120. Let each answer be, on average 2 lines. Including reading and evaluation, this should be about 1 minute of "live" time. Including rest time, 1 coder can process 50 questions in 1 working hour, which means 400 questions in 1 working day (8 hours per day), which is 200 working days for 80,000 students. This means that if we calculate a relatively cheap daily rate (let us say 10,000 HUF, for which we might not be able to find a coder, but let us say we can), coding one question costs 20,000,000, or 20 million HUF. Of course, it does not always take 1 minute to code an answer to such a question, but even if we calculate the time in 10 seconds, we will reach one-sixth of the cost, i.e., in the order of 100 million HUF. It is an understandable suggestion on the part of the commissioner to investigate, whether these costs can be reduced by omitting open-ended tasks; or by formulating them as a closed task. From a professional point of view, it is reasonable to doubt whether the omission of open-ended items results in the same measurement.

Several questions may arise from this thought experiment:

1. What area is measured by open-ended questions (Bridgeman, 1992; Geer, 1991)?
2. Is it important to measure these areas (Groves, 1978)?
3. Can open-ended questions be replaced by questions that can be automatically scored (Reja et al., 2003)?
4. Do all subjects need open-ended questions (Eilam, 2002)?

The basic question of our study – although the others are also valid – is the fourth one. The first two questions are professional questions: which areas are important to measure and in what quality? The third question in our view is more of a technological question. Here we list, for example, the development of innovative items, like problem-solving and inquiry tasks simulating real-life and laboratory situations (Mullis & Martin, 2017) and computer-supported coding systems that take advantage of the possibilities of computerized measurement, e.g., the machine-supported coding system, which was developed for PISA 2015 (Yamamoto et al., 2017). We want to investigate the fourth question in more detail in this study.

Open-ended questions (e.g., projective tests) can also be used in labour market surveys, which can be taken using computers, but in which the practitioner may play the primary role (Darby, 2007). Our research question is not whether it is worth asking open-ended questions in a labour market selection situation but whether it is worth asking them from everyone (Metzner & Mann, 1952). Even more importantly, is it in the client's and employer's interest to ask questions regarding live expertise from all candidates? The answer to this question is clear: of course not (Raub & Streit, 2006). The answer to the first question is negative because logically we do not want to measure every applicant according to every aspect. Let's think about the situation that if the application requires a driver's license, we do not want to interview and talk to applicants who do not have it. And similarly: clients do not apply for job offers where they are expected to have qualifications that they do not possess. If only people with a medical degree can apply for a job, people without a medical degree won't submit their resumes – or they won't expect to be called in for an interview. This leads to the second question: to whom should we ask these questions?

Previous studies (Izsó, Berényi & Takács, this special issue) have shown that there is a significantly higher probability than a chance of selecting subjects for whom more costly but more nuanced questions are justified. However, the other half of the questions should not be bypassed. What information is lost for those for whom these questions should have been asked but were not? The heigh end of the ability range, where the open-ended question system can provide additional information, is typically the category of those who perform very well. In a labour market situation, however, their less accurate knowledge is not an actual loss – since they are the ones who are typically invited to a recruitment interview as a result of automatic selection, and in their case, we ultimately use live expertise, so there is no actual loss.

The lower end of the ability range in a labour market typically represents the "very poor performers". Open questions during pre-screening calls provide additional information for candidates who would usually not be invited for an interview. In their case, the automated tests will show that they are not good candidates, but the questions to be coded will give us a better understanding of why they are not good candidates.

In another labour market situation, employers monitor their employees for prevention or development purposes. It can be a matter of preventing turnover, training, or maintaining mental health, skills development, skill-based integration, a more precise exploration of integration into a collective, or simply the clarification and better mapping of integration. In a labour market measurement setting, the function of open-ended tasks in the lower region of an ability scale can be, for example, clarifying and understanding the areas to be developed, and finding the deeper reasons for uncovering blockages. All in all, automated questions can help identify those who need to be targeted by professionals – because they are the ones who need help, even at the individual level, and it is the low performers who will be more closely screened.

Measuring competences

The measurement of competencies has already been discussed in several places inthis thematic issue (Izsó, Berényi & Takács, this special issue; Pusker, Gergely & Takács, this special issue), so in this paper, we will only cover the area that is necessary to interpret the results of the calculation. In our article, we have used the item-level results of a large-scale measurement to explore the implications of omitting open-ended questions for larger measurement systems.

National Assessment of Basic Competencies

Several studies on the National Assessment of Basic Competencies have been published in the last 20 years since it was organised annually in Hungary. The measurement results can be found in the national reports (Belinszki et al., 2020). Due to the large volume of the measurement and the broad spectrum covered by the background questionnaires, it also serves as a source of data for several secondary analyses (Kövesdi et al., 2020; Nyitrai et al., 2020; Szemerszki, 2015).

In the case of the National Assessment of Basic Competencies, there is no declared content domain or thinking operation for open or closed questions, which means that open-ended questions can be used in either reading comprehension or mathematical competencies. There is no specific operational or competence domain division that requires the use of open-ended questions (Balázsi et al., 2014).

We note that out of the 4 questions we asked earlier, these documents also answer the first two questions proposed. The assessment organizer's surveys also showed that open-ended questions are not necessarily justified in all areas – and there is no feedback reported on each area separately. However, this does not mean that it is not possible to formulate the expectations in different content domains at a given proficiency level. In the case of students performing at a certain level, it can be clearly stated what kind of solution we can expect from them in a specific type of task, in what quality they can solve the problems in the predetermined area. But this also means that the measuring organization dealt with serious dilemmas until they were able to make this statement. It seems legitimate that such a decision of a certain company (in which part of the selection or evaluation process would open-ended questions be important) should be preceded by the same discussion.

In the area of reading, the thinking operations are as follows (Balkányi et al., 2018):

1. Information retrieval;
2. Recognizing connections and relationships;
3. Interpretation.

The same in the area of mathematics (Lak et al., 2018):

1. Fact recognition and simple operations;
2. Application and integration;
3. Complex solutions and evaluation.

These operations by mathematical tools are used in tasks measuring the following content areas:

A. Quantity, numbers, operations;
B. Assignments, relationships;
C. Shapes, orientation;
D. Statistical properties, probability.

After coding the tasks and calculating the scores, an IRT model is used to calculate both the difficulty of the tasks and the students' performance (Auxné Bánfi et al., 2014). Then, for easier understanding and interpretation, seven ability levels are set both in reading comprehension and mathematics. Expected performances and skills are assigned based on the types of tasks corresponding to the levels of difficulty and the thinking operations required for them.

Based on the content framework, tasks are sorted into test booklets according to thinking operations and content area/text type (Balázsi et al., 2014). According to the task format, open-ended coding questions requiring longer answers are assumed to be among the more complex tasks. Thus, their real informational contribution appears in the "higher performance regions".

Knowledge and skills

At this point, it is worth identifying the areas of competencies we are discussing. Some areas can be achieved, for example, through studying, retrieval, and memorization of information, and these are called knowledge (Eraut et al., 2000). Automated items can measure this area quite well (National Research Council, 2012).

In contrast, there are domains, which are more of a practical expertise (Spenner, 1990). For example, knowledge is similar to an exam regarding traffic regulations where one knows the right answer to a question (one must slow down and give priority at a priority sign) – while in the case of skills, considering a real-life scenario while driving in traffic one actually slows down and give priority. All this does not mean that automated items cannot measure domains, but they may require more preparation or measurement tools in some workplace settings. One such measurement tool is the ErgoScope (Izsó, Berényi & Pusker, this special issue), which can be considered an automated assessment in that a machine automatically provides the data. However, it is still a "live" measurement, where a trained assistant is needed to operate the machine, so its use may require considerable resources on the client's part. In this sense, ErgoScope is more in the category of "open" questions. The use of the measurement tool is reflected at length in the ErgoScope study (Izsó, Berényi & Pusker, this special issue), where the other extreme of the recruitment narrative for proficiency is explored, the reasons for low performance. In particular, in the case of the "under-performers" mentioned earlier, we see added value in terms of what barriers, such as physical performance, may impede the worker's potential placement.

Automatic evaluation

By automatic scoring, we mean a system like the one described in the study by Gergely and Takács (this special issue). In such system, a computer provides the questions to the subjects and offers the expert with aggregated results from the answers received. By expert, we mean an HR staff member, a support professional or a teacher. The point is that it is not the expert who evaluates the results of the questionnaire survey (or even a school essay) but works with aggregated results.

In the case of ATOM (Gergely & Takács, this special issue; Izsó, Berényi & Pusker, this special issue), this may even mean evaluating individual elements of CVs, thus facilitating the collection and evaluation of information on the minimum requirements for a given job.

The time gained through the evaluation can then be used by the professional to address questions and areas that computerised evaluation systems are currently unable to address or are very limited in their ability to do, for example:

1. After the evaluation, the teacher can investigate the possible shortcomings behind the failed tasks. Of course, "skilful guessers" in closed questions can remain hidden but let us assume that in the mass of automatically scored tasks, simple guessers cannot answer all questions correctly (Brassil & Couch, 2019).
2. In an ErgoScope-type test, there may be several physical or other deficiencies behind the errors or underperformance. A face-to-face discussion with a specialist can help to identify the reasons (Izsó, Berényi & Pusker, this special issue).
3. The HR representative usually does not invite all candidates to the interview but only potential candidates who meet the eligibility criteria. At the same time, the pre-assessment of the candidates is carried out by an automatic evaluation system, which frees up time for the HR professional to interview several potentially suitable candidates in person within the same time limit. It should be noted here that the automatic assessment system (Izsó, Berényi & Pusker, this special issue) can send essentially personalised feedback to all candidates, so that even those candidates who are not ultimately met in person by HR staff (Izsó, Berényi & Pusker, this special issue) will receive some form of personalised message.

Thus, automatic assessment systems are expected to support the work of professionals so that a more significant proportion of professional time can be devoted to working processes requiring expertise (Fawcett, 1992).

Continuous or categorical feedback

The form of feedback is a methodologically important issue since it makes a difference whether the predictive outcome indicates a continuous indicator of achievement (e.g., a percentage achievement) or a categorical indicator of achievement (Gergely & Takács, this special issue; Izsó, Berényi & Pusker, this special issue). In the case of the National Assessment of Basic Competencies, the performance variable indicates a continuous indicator of achievement. At the same time, the National Assessment of Basic Competencies, like other international measures of student performance, maps performance to so-called achievement levels (e.g., OECD PISA [OECD, 2019]).

The performance levels obtained at the end of the assessment overlap significantly with the interpretation of the categories of entry into the workplace since the interpretation of the categories and levels obtained in the competency assessment implies a kind of "expected knowledge, provided knowledge". It shows us what tasks a student at a given level is most likely to be able to perform independently and confidently (Balázsi et al., 2014).

This approach is methodologically equivalent to the categories of workplace validation. The assessment of workplace compliance (eligible/not eligible, or level of compliance) also carries a similar meaning. In our view, the analysis of the National Assessment of Basic Competencies' student performance can be well applied to our evaluation system, as these evaluation systems are similar in several respects:

1. Students are not assessed by their teachers but are assessed using an external measurement tool (see ErgoScope's measurement technology [Izsó, Berényi & Pusker, this special issue]).
2. Students' performance is measured on a continuum of scales and then categorised into performance levels (Izsó, Berényi & Takács, this special issue).
3. A large amount of measured data is available to visualise shifts at a mass level, not just individual cases (Gergely & Takács, this special issue).

The National Assessment of Basic Competencies was implemented in digital format for the first time in 2022 after 20 years of paper and pencil testing (Oktatási Hivatal, 2021), so the issue of automated assessment is also current.

Based on this, our hypotheses are:

1. The performance computed from closed items with automatic coding is a good approximation of the performance computed from automatic and live coding. We expect that the ability scores computed in the two ways should show correlations around 0.9. This means, in simple terms, that although we assume differences between the scores without full and open questions, the questions and tasks capture the same domain at the substantive level.
2. At the lower levels, we typically see an "upward" bias (namely: without open-ended questions, students perform essentially "better"). On the labour market side, this suggests that those with typically lower labour market status are better off when evaluated with closed items (Podsakoff & Organ, 1986).
3. At higher levels, the opposite is expected (good answers to open-ended items typically make good students look "even better"). By omitting open-ended items, workers with typically good labour market status are less able to stand out, somewhat "blending in" with their environment. In their case, a personal interview, for example, may be necessary to refine the selection (Vázquez-Alonso et al., 2006).

Sample and methodology

The results of the student-level data are presented from the main survey in 2017 at the 6th-grade level. In the measurement 91,599 students participated who were required to take the measurement, of which 85,563 students had a completed test booklet and an assessable score after absences and total exemptions. However, not all of these students were eligible (e.g., some students with special educational needs are not exempted from participation, but their results are not included in the aggregated results), so ultimately, 81,647 students' data remained after excluding those with exemptions from the complete analysis.

A specific feature of the National Assessment of Basic Competencies is that it essentially measures the current population (Belinszki et al., 2020), i.e., the sample can be considered representative of this stratum. Therefore, weighting was applied following the methodology of the National Assessment of Basic Competencies (Auxné Bánfi et al., 2014) so that the results are representative of a total of 86,151 students. Ability scores from closed items were calculated using the Parscale 4.1 software package, and further calculations were performed using the IBM SPSS 28.0 software package.

The tests were performed at a 95% significance level. Pearson correlation was used to test the relationship between ability scores. For the cross-tabulation analyses, the significance of the chi-squared test and the adjusted standardised residuals were included as effect sizes by category.

Methodological overview

There are participants from 3 different grades in the NABC (6th, 8th, and 10th grades). Grade 8th data are typically included after admission and once the results are known. The motivational background may be questionable in general cases, but this may be more pronounced here for grade 8th. Grade 10th produces the *"better"* results for the whole population, but this would "present" a labour market situation where we are in the fortunate position of typically having the *"best"* candidates for an advertised job. Since we do not focus on this labour market situation but rather on a situation where selection can be interpreted as a natural, genuine selection process. This type of selection of the cohort was of no material relevance for the interpretation of the results. Of the 3 possible age groups, tables from grade 6th are in the main text, and the other 2 groups' results are in the appendix.

Results

Item-level data were used to calculate two types of scores per student: on the one hand, using performance scores from the entire test (with both open-ended and closed items), and on the other hand, using performance scores from a "shorter" test consisting of only closed items. That is: for each student, we have a score where his/her open answers are coded and one where we have asked the scoring system to "automatically evaluate".

We will first look at the coincidences for the continuous outcomes and then at the coincidences for the categorisation.

Correlation coefficients – covariance of continuous scoring

In the first step, we examined the Pearson correlation between the scores calculated from the full test and the "closed only" questions *(Table 1).* On the Pearson correlation coefficients, we observe that the correlation coefficients are sufficiently high for measures in the same domains. From this comes that the reading comprehension and mathematics scores are correlated with each other at the expected level of between 0.7 and 0.8. In contrast, the scores from the closed questions show a correlation with the corresponding entire test scores above 0.9.

Table 1. Correlations of ability scores calculated on the entire test and closed items only

Pearson correlation coefficients **N_6 = 86151** **N_8 = 80833** **N_{10} = 76550**		**Math Score, FULL TEST**	**Reading Score, FULL TEST**	**Math Score, CLOSED**	**Reading Score, CLOSED**
Math Score, FULL TEST	6th grade 8th grade 10th grade	–	.723** .777** .775**	.910** .954** .963**	.703** .741** .752**
Reading Score, FULL TEST	6th grade 8th grade 10th grade	.723** .777** .775**	–	.674** .739** .741**	.932** .958** .951**
Math Score, CLOSED	6th grade 8th grade 10th grade	.910** .954** .963**	.674** .739** .741**	–	.664** .716** .729**
Reading Score, CLOSED	6th grade 8th grade 10th grade	.703** .741** .752**	.932** .958** .951**	.664** .716** .729**	–

Note: **: p < 0.01

Cross tabulation analyses

We then compared the levels resulting from the two scores in mathematics and reading comprehension to see in which directions the variance of the scores is skewed when looking at the bigger picture. This kind of "individual" variation is nuanced by trying to capture the level of students' scores rather than their scores. National Assessment of Basic Competencies' ability scale is constructed with a mean of 1,500 points and a standard deviation of 200 points, suggesting a possible range of scores between 1,200 and 1,800. The competency scale is divided into 8 levels, with a "score width" of approximately 100 points per level. In addition, the standard error of students' performance is of 50–80 points, so we can expect a change in ability level if the score is on the "borderline" of two levels.

Adjusted residuals (AR) for cross tables indicate that the number of observed cases in the given cell is lower (negative AR) or higher (positive AR) than expected number in the case of independence. Values greater than 2 or less than -2 already indicate a difference. It can be observed in the case of all three grades that in the higher levels, both types of distortion typically occur with the omission of open-ended items (for the 6th grade, see *Table 2,* for the 8th grade and 10th grade, see *Appendix 1* and *Appendix 2*). This ratio is about 20% in both upward and downward distortion. In the lower regions, test subjects typically perform better by omitting open-ended questions. About 40% of the true "Below 1st level" and 1/3 of the true "1st level" students categorized to the next proficiency level.

In other words, better performing students display better results on the typically harder, open-ended questions. As a consequence, however, the need for open questions arises already at proficiency levels 5 and 6, i.e., slightly above the average level of proficiency in mathematics. This difference-in-difference means the following: given two students whose mathematics performance is examined for total scores and closed questions. If student A performs better than student B on the total measure, then student A cannot maintain the "leading role" by omitting open questions (yellow background), or at least, there is uncertainty in classifying. However, even more striking is that the test subjects in the lower levels are valued upwards by the lack of open questions (yellow background). In other words, those with a lower real performance appear in a better light by omitting the open-ended questions.

Table 2. Comparison of ability levels in 6th grade between the full test and the closed items only in mathematics. If the expected count is less than the observed count, one level distortion from the correct class towards the center is marked with yellow background, towards the extremes is marked with green background

Below 1st **1st**			**Math Proficiency Level, CLOSED**								**Total**
			2nd	**3rd**	**4th**	**5th**	**6th**	**7th**			
Math Proficiency Level, FULL TEST	Below 1st	Count	**1936**	1554	2	0	0	0	0	0	3492
		AR	**191,2**	64,6	-33,0	-37,1	-30,5	-20,8	-11,6	-6,0	
	1st	Count	486	**6051**	2994	28	0	0	0	0	9559
		AR	14,0	**173,3**	20,1	-63,1	-52,5	-35,8	-19,9	-10,3	
	2nd	Count	21	1829	**13557**	4118	46	0	0	0	19571
		AR	-26,2	-8,4	**174,1**	-22,9	-79,6	-54,9	-30,5	-15,9	
	3rd	Count	0	43	3326	**16641**	4505	132	0	0	24647
		AR	-31,7	-64,3	-42,5	**166,8**	-9,8	-60,9	-35,6	-18,5	
	4th	Count	0	0	54	2852	**11521**	3408	180	24	18039
		AR	-25,8	-53,1	-81,8	-39,4	**162,9**	40,4	-20,8	-13,0	
	5th	Count	0	0	0	17	1504	**5165**	1391	147	8224
		AR	-16,3	-33,5	-52,3	-58,2	-5,0	**161,3**	69,0	7,8	
	6th	Count	0	0	0	0	4	462	**1371**	368	2205
		AR	-8,1	-16,7	-26,1	-29,3	-23,9	15,9	**151,0**	75,7	
	7th	Count	0	0	0	0	0	0	106	**308**	414
		AR	-3,5	-7,2	-11,2	-12,5	-10,3	-7,0	24,4	**151,8**	
Total		Count	2443	9477	19933	23656	17580	9167	3048	847	86151

Note: Count is Observed Frequencies and AR is Adjusted Standardized Residual

In the case of reading comprehension, the role of open questions is less critical, but the situation shows similar dynamics (for the 6th grade, see *Table 3,* for the 8th grade and

10th grade, see *Appendix 3* and *Appendix 4*). In the lower region, there is a greater bias in the direction of better abilities, while in the case of the upper regions, downward bias will continue to be more typical (both marked with yellow background). We can also say that the bias appears later in the case of reading comprehension – if you like, we can measure a larger range of ability levels with closed items at an acceptable level.

Table 3. Comparison of ability levels on the 6th grade between the entire test and the closed items only in reading. If the expected count is less than the observed count, one level distortion from the correct class towards the center is marked with yellow background, towards the extremes is marked with green background.

Below 1st 1st			**Reading Proficiency Level, CLOSED**								**Total**
			2nd	**3rd**	**4th**	**5th**	**6th**	**7th**			
Reading Profici-ency Level, FULL TEST	Below 1st	Count	**760**	503	0	0	0	0	0	0	1263
		AR	**202,5**	50,0	-15,0	-19,5	-21,2	-17,2	-10,7	-5,1	
	1st	Count	188	**4017**	1413	0	0	0	0	0	5618
		AR	16,7	**210,3**	22,0	-42,2	-45,9	-37,2	-23,2	-11,0	
	2nd	Count	2	804	**9676**	2482	10	0	0	0	12974
		AR	-12,9	0,1	**206,1**	-11,0	-73,0	-59,3	-36,9	-17,6	
	3rd	Count	0	0	1840	**14716**	3615	101	5	0	20277
		AR	-17,2	-41,8	-27,1	**192,8**	-30,2	-76,0	-48,5	-23,2	
	4th	Count	0	0	0	2499	**16216**	3464	123	8	22310
		AR	-18,3	-44,5	-72,9	-48,2	**185,0**	-14,1	-48,4	-24,3	
	5th	Count	0	0	0	0	2520	**11354**	2061	93	16028
		AR	-14,8	-36,0	-59,0	-76,4	-32,8	**187,6**	23,8	-14,1	
	6th	Count	0	0	0	0	0	1197	**4454**	750	6401
		AR	-8,8	-21,3	-34,9	-45,3	-49,2	0,0	**186,1**	58,0	
	7th	Count	0	0	0	0	0	0	420	**860**	1280
		AR	-3,8	-9,3	-15,1	-19,6	-21,3	-17,3	32,3	**168,5**	
Total		Count	950	5324	12929	19697	22361	16116	7063	1711	86151

Note: Count is Observed Frequencies and AR is Adjusted Standardized Residual

For the cross-tables, we expect to find large values in the "main diagonal" cells (with a gray background and bold typeface) connecting the North-West corner with the South-East corner and fewer cases as we move away from there. Furthermore, at any level, the deviation of the closed items from the entire test score by 2 levels is very rare (less than 1%). We have seen this in both cases: for math, we see an "upward" bias in the lower region (downward bias in the upper), and for reading comprehension, the more significant biases tended to be at the higher levels.

Discussion

The National Assessment of Basic Competencies data allowed us to test approximately 3 × 80,000 respondents the mass consequences of omitting open-ended items that may have been crucial for our project. In the case of the National Assessment of Basic Competencies, the question arose as to what justifies the use of different open-ended tasks and when. They came to the conclusion that in this area it is not necessary to use open-ended items for more accurate measurement of content domains or thinking operations – but it certainly makes the survey as a whole more colourful and varied (Balázsi et al., 2014). However, this cannot always be said for a workplace selection process since open-ended questions and an interview with the future superior remain an essential part of the process. Our question was not whether it is possible to remove the entire process. Our question was more about when or at which point should they be used in the entire process.

We did not consider the level of individual feedback important- as this role is reserved for professionals in our testing situation (Izsó, Berényi & Pusker, this special issue).

We primarily addressed the question of what biases might be expected in a larger-scale application of an automatic evaluation system by omitting open-ended questions (if you like, by reclassifying the live evaluation) (Brassil & Couch, 2019; Bridgeman, 1992). In our study, we wanted to test whether using only closed items rather than a combination of open- and closed questions would result in the same decisions when categorizing respondents.

Our calculations demonstrated that we could detect a reasonably close correlation level above 0.9 between the ability scores calculated using the entire test and the closed item only scores in the continuous evaluations. It is crucial to note the condition that the National Assessment of Basic Competencies is based on relatively high-quality and multiple-tested questions (Auxné Bánfi et al., 2014), which also guarantees the omission of some questions does not cause system-level problems. This last result is perhaps the most important: it means that calibration, the classification into levels using closed questions, does not lead to a misclassification of more than 2 levels for 8 ability levels! It also means that omitting open-ended items does not generate a bias greater than the width of a level, with a shift of more than twice the standard error essentially undetectable by omitting open-ended questions.

This is obviously a limitation of our study: The National Assessment of Basic Competencies is a comprehensive survey, so we have accurate aggregated data at the national and regional level. This is not the case of a workplace recruitment. In the case of competence measurement, we may identify possible development areas for students, or provide feedback to the teacher about the competence level of the classes in comparison to other student groups or classes, so such assessments need to survey test subjects with the same precision. In the case of workplace selection, however, the primary aim is the selection of the best applicant(s). There is also another type of limitation: in a workplace situation, the applicant has a serious stake in responding. This is not the same in the case of the National Assessment of Basic Competencies: typically, this is a low-stake test for the students (at least we cannot talk about a stake situation from the side of the students in relation to the Educational Authority, who conducting the survey) (Auxné Bánfi et al., 2014). In a selection situation – such as a high school or a university admission procedure – the admission committees should not devote significant resources to the most unsuitable candidates. This means that with the help of the automated item lines, it is possible to outline those candidates with whom we really want to conduct longer, more resource-intensive examinations. Of course, this selection can also aim for development in the workplace, or also for a talent management.

However, in our opinion, the following analogy stands firm even with this limitation and difference. The proficiency levels of the National Assessment of Basic Competencies include expected achievement, based on which it can be said that the student at a given level is capable of solving tasks in a subject area. This type of classification can be considered analogous to the procedure of workplace selection. In this sense, with the examination of advantages and disadvantages regarding the use of automation while applicant classification appears to be an analogous problem. So, the phenomena experienced here also serve as a reference point during workplace selection.

Category-level analyses of the results showed that there were typically significant differences at the two extremes of our measurement scale, which is consistent with the results of Geer and colleagues (1991). While those who performed at the lower levels seemed to have a slightly better performance, in the case of those who performed at the higher levels, less uncertainty can be observed. In the middle performance range, the two types of test results led to a similar classification. It seems reasonable to apply open-ended questions (the evaluation of which is more costly and complicated than the evaluation of items that can be automated) only to who performed in the upper (or in the case of development, lower) levels on the closed questions test. This only partially coincides with the previous result of Balázsi et al. (2014), since they did not find a measurement reason for the application in any content area. However, according to our hypothesis, we found that after an automated classification, it is indeed worthwhile to use open-ended items for candidates on the upper levels - however, this does not mean that we have to ask all applicants these questions in a selection process.

Our experience and calculations show that the involvement of professionals in the selection process can be delayed until later, in the sense that they are more likely to have to conduct personal interviews with suitable candidates. In conclusion, we see that the role of professionals cannot be neglected in the selection process (Izsó, Berényi & Pusker, this special issue; Izsó, Berényi & Takács, this special issue), nor can the expertise of teachers be neglected in classroom assessment.

We also highlight that closed items in the lower regions of the performance scales were associated with the opposite bias. This implies that the practitioner can use the face-to-face assessment to uncover hidden problems, the longer-term concealment of which may be associated with health problems for the subjects. In the longer term, ErgoScope examinations may be more important in preventing staff turnover and safeguarding workers' health (Izsó, Berényi & Pusker, this special issue).

Összefoglaló

Nyílt és zárt itemek használata kiértékelési rendszerek automatizálásában

Háttér és célkitűzések: Automatizált kiértékelési rendszerek tervezésének során felmerül a nyílt végű kérdések, avagy az humán szakértelmet kívánó feladatok elhagyásának problémája. A nyílt végű kérdések kódolása költséges, idő és munkaerőigényes feladat. Az életrajzok átnézése és kiválogatása csökkenti a személyes interjúkra fordítható időmennyiséget. Kutatási kérdésünk ennek mentén az, hogy az ilyen szempontok elhagyása mely tesztalanyok esetében és milyen következménnyel jár az értékelés eredményét tekintve. Vizsgálatunkat az *Országos kompetenciamérés* adatain végeztük, amely önmagában szintén felfogható egy értékelő rendszerként, és amely jelenleg az automatizált kiértékelés bevezetésének fázisában van.
Módszer: Az eredeti arányokat a mérés módszertana szerinti súlyozással állítottuk vissza. Vizsgálatunkban összehasonlítottuk a teljes tesztfüzet alapján és a kizárólag zárt itemek alapján számított teljesítménypontokat és képességszinteket.
Eredmények: A teljes tesztből és a csak zárt itemekből számított képességpontok igen erős összefüggést mutatnak.
Következtetések: Számításaink azt igazolják, hogy a nyílt végű itemekre azokban a képességtartományokban van szükség, ahol eleve kevesebb item áll rendelkezésre. A nyílt végű kérdések elhagyásával nagy „veszteség” jellemzően azokat éri, akikről kevesebb információval rendelkezünk, akiket a „nagyon jól teljesítő” és a „nagyon rosszul teljesítő” kategóriákba sorolunk.
Kulcsszavak: automatizált kiértékelés, munkahelyi beválás, értékelési rendszer, *Országos kompetenciamérés*

References of this Special Issue

Gergely, B., & Takács, Sz. (2023). ATOM – a flexible multi-method machine learning framework for predicting occupational success. *Alkalmazott Pszichológia*, *25*(3), 15–30.

Izsó, L., Berényi, B., & Pusker, M. (2023). Jointly applying a work simulator and ATOM to prevent occupational accidents and MSD through workforce selection. *Alkalmazott Pszichológia*, *25*(3), 73–91.

Izsó, L., Berényi, B., & Takács, Sz. (2023). Illustrating real-life ATOM application case studies. *Alkalmazott Pszichológia*, *25*(3), 93–114.

Pusker, M., Gergely, B., & Takács, Sz. (2023). ATOM's structure – employee and employer feedback, survey site. *Alkalmazott Pszichológia*, *25*(3), 53–72.

References

Auxné Bánfi, I., Balázsi, I., Balkányi, P., Balogh, V. K., Gyapay, J., Lak, Á. R., Ostorics, L. I., Palincsár, I., Rábainé Szabó, A., Rózsa, Cs., Szabó, Á., Szabó, L. D., Szepesi, I., Szipőcsné Krolopp, J., & Vadász, Cs. (2014). *Országos kompetenciamérés, Technikai leírás*. Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2012/OKM_Technikaileiras.pdf

Balázsi, I., Balkányi, P., Ostorics, L., Palincsár, I., Rábainé Szabó, A., Szepesi, I., Szipőcsné Krolopp, J., & Vadász, Cs. (2014). *Az Országos kompetenciamérés tartalmi keretei – Szövegértés, matematika, háttérkérdőívek*. Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2014/AzOKMtartalmikeretei.pdf

Balkányi, P., Gyapay, J., Lak, Á. R., Szabó Rábainé, A., Suhajda, E., Szabó, L. D., & Takácsné Kárász, J. (2018). *Országos kompetenciamérés 2017. Feladatok és jellemzőik szövegértés 6. Évfolyam*. Oktatási Hivatal, Köznevelési Mérés Értékelési Osztály. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2017/OKM2017_Feladatok_es_jellemzoik_Szovegertes_6.pdf

Belinszki, B., Szepesi, I., Takácsné Kárász, J. & Vadász, Cs. (2020). *Országos jelentés 2019. Országos kompetenciamérés*. Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2019/Orszagos_jelentes_2019.pdf

Bingölbali, E., & Bingölbali, F. (2021). An Examination of Open-Ended Mathematics Questions' Affordances. *International Journal of Progressive Education*, *17*(4), 1–16. https://doi.org/10.29329/ijpe.2021.366.1

Brassil, C. E., & Couch, B. A. (2019). Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: A Bayesian item response model comparison. *International Journal of STEM Education*, *6*(16), 1–17. https://doi.org/10.1186/s40594-019-0169-0

Bridgeman, B. (1992). A Comparison of Quantitative Questions in Open-Ended and Multiple-Choice Formats. *Journal of Educational Measurement*, *29*(3), 253–271. https://doi.org/10.1111/j.1745-3984.1992.tb00377.x

Çınar, A., Ince, E., Gezer, M., & Yılmaz, Ö. (2020). Machine learning algorithm for grading open-ended physics questions in Turkish. *Education and Information Technologies*, *25*(5), 3821–3844. https://doi.org/10.1007/s10639-020-10128-0

Darby, J. A. (2007). Open-ended course evaluations: A response rate problem? *Journal of European Industrial Training*, *31*(5), 402–412. https://doi.org/10.1108/03090590710756828

Eilam, B. (2002). Strata of comprehending ecology: Looking through the prism of feeding relations. *Science Education*, *86*(5), 645–671. https://doi.org/10.1002/sce.10041

Eraut, M., Alderton, J., Cole, G., & Senker, P. (2000). Development of knowledge and skills at work. In Coffield, F. (Ed.), *Differing visions of a learning society: Research findings*. 1. Policy Press, Bristol. 231–262. https://doi.org/10.56687/9781847425126-009

Fawcett, W. (1992). Staff satisfaction in new offices: Findings of an interactive computer questionnaire. *Property Management*, *10*(4), 338–347. https://doi.org/10.1108/02637479210030475

Geer, J., G. (1991). Do Open-ended questions measure ‘salient’ issues? *Public Opinion Quarterly*, *55*(3), 360–370. https://doi.org/10.1086/269268

Groves, R. M. (1978). On the mode of administering a questionnaire and responses to open-ended items. *Social Science Research*, *7*(3), 257–271. https://doi.org/10.1016/0049-089X(78)90013-3

Kövesdi, A., Kovács, D., Harsányi, Sz. G., Koltói L., Nagybányai-Nagy, O., Nyitrai, E., Simon, G., Takács, N., & Takács, Sz. (2019). A 2018. évi Országos kompetenciamérés eredményei Magyarországon – Az SNI-vel és BTM-mel diagnosztizált 6., 8., 10. évfolyamos gyermekek körében. *Psychologia Hungarica Caroliensis*, *7*(4), 52–122.

Lak, Á. R., Palincsár, I., Szabó, L. D., Szepesi, I., Szipőcsné Krolopp, J., & Takácsné Kárász, J. (2018). *Országos kompetenciamérés 2017. Feladatok és jellemzőik matematika 6. évfolyam*. Oktatási Hivatal, Köznevelési Mérés Értékelési Osztály. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2017/OKM2017_Feladatok_es_jellemzoik_Matematika_6.pdf

Lafontaine, D., & Monseur, C. (2009). Gender Gap in Comparative Studies of Reading Comprehension: To What Extent Do the Test Characteristics Make a Difference? *European Educational Research Journal*, *8*(1), 69–79. https://doi.org/10.2304/eerj.2009.8.1.69

Metzner, H., & Mann, F. (1952). A Limited Comparison of two Methods of Data Collection: The Fixed Alternative Questionnaire and the Open-Ended Interview. *American Sociological Review*, *17*(4), 486–491. https://doi.org/10.2307/2088007

Mullis, I. V. S., & Martin, M. O. (Eds.) (2017). *TIMSS 2019 Assessment Frameworks*. TIMSS & PIRLS.

National Research Council (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century* (Pellegrino, J. W. & Hilton, M. L., Eds.). The National Academies Press.

Nyitrai, E., Harsányi, Sz. G., Koltói, L., Kovács, D., Kövesdi, A., Mátai, G.., Nagybányai-Nagy, O., Pusker, M., Simon, G., Smohai, M., Takács, N., & Takács, Sz. (2019): Szülői bevonódás és az iskolai teljesítmény kapcsolata az Országos kompetenciamérés 2017-es és 2018-as adatainak tükrében. *Psychologia Hungarica Caroliensis*, *7*(4), 7–51.

OECD (2019). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing, Paris. https://doi.org/10.1787/b25efab8-en

Oktatási Hivatal (2021). *A digitális országos mérések általános leírása*. https://www.oktatas.hu/kozneveles/meresek/digitalis_orszagos_meresek/altalanos_leiras

Podsakoff, P. M., & Organ, D. W. (1986). Self-Reports in Organizational Research: Problems and Prospects. *Journal of Management*, *12*(4), 531–544. https://doi.org/10.1177/014920638601200408

Raub, S., & Streit, E. M. (2006): Realistic recruitment: An empirical study of the cruise industry. *International Journal of Contemporary Hospitality Management*, *18*(4), 278–289. https://doi.org/10.1108/09596110610665294

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003): Open-ended vs. Close-ended Questions in Web Questionnaires. *Developments in Applied Statistics*, *19*(1), 159–177.

Spenner, K. I. (1990). Skill: Meanings, Methods, and Measures. *Work and Occupations*, *17*(4), 399–421. https://doi.org/10.1177/0730888490017004002

Szemerszki, M. (2015). A tanulói teljesítménymérések szerepe a tényekre alapozott oktatáspolitikában. In Széll K. (Ed.), *Mit mér a műszer?* (pp. 9–22). Oktatáskutató és Fejlesztő Intézet.

Vázquez-Alonso, Á., Manassero-Mas, M.-A., & Acevedo-Díaz, J.-A. (2006). An analysis of complex multiple-choice science–technology–society items: Methodological development and preliminary results. *Science Education*, *90*(4), 681–706. https://doi.org/10.1002/sce.20134

Yamamoto, K., He, Q., Shin, H. J., & von Davier, M. (2017). *Developing a Machine-Supported Coding System for Constructed-Response Items in PISA*. Research Report. ETS RR-17-47. ETS Research Report Series. https://files.eric.ed.gov/fulltext/EJ1168681.pdf https://doi.org/10.1002/ets2.12169

Appendices

Appendix 1. Comparison of ability levels on the 8th grade between the full test and the closed items only in mathematics

Below 1st **1st**			**Math Proficiency Level, CLOSED**								**Total**
			2nd	**3rd**	**4th**	**5th**	**6th**	**7th**			
Math Profici-ency Level, FULL TEST	Below 1st	Count	**828**	704	0	0	0	0	0	0	1532
		AR	**180,1**	65,4	-16,0	-21,1	-22,7	-19,0	-12,7	-7,2	
	1st	Count	250	**2977**	1387	8	0	0	0	0	4622
		AR	24,6	**169,7**	32,0	-37,1	-40,2	-33,7	-22,4	-12,8	
	2nd	Count	17	1243	**7274**	2115	28	0	0	0	10677
		AR	-11,5	25,3	**172,1**	-6,4	-63,0	-53,4	-35,5	-20,3	
	3rd	Count	0	51	2685	**11893**	3271	74	3	0	17977
		AR	-17,8	-37,1	3,6	**160,9**	-23,3	-71,6	-48,6	-27,8	
	4th	Count	0	0	62	3881	**13509**	3366	109	6	20933
		AR	-19,7	-43,0	-66,7	-14,8	**154,6**	-11,8	-50,8	-30,5	
	5th	Count	0	0	0	43	3240	**10124**	2280	107	15794
		AR	-16,4	-35,9	-56,8	-73,9	-14,0	**162,3**	24,7	-20,3	
	6th	Count	0	0	0	0	11	1654	**4613**	883	7161
		AR	-10,4	-22,7	-35,9	-47,3	-50,6	9,7	**168,2**	45,6	
	7th	Count	0	0	0	0	0	0	517	**1620**	2137
		AR	-5,5	-12,0	-19,0	-25,0	-26,9	-22,6	24,0	**192,1**	
Total		Count	1095	4975	11408	17940	20059	15218	7522	2616	80833

Note: Count is Observed Frequencies and AR is Adjusted Standardized Residual

Appendix 2. Comparison of ability levels on the 10th grade between the full test and the closed items only in mathematics

Below 1st **1st**			**Math Proficiency Level, CLOSED**								**Total**
			2nd	**3rd**	**4th**	**5th**	**6th**	**7th**			
Math Profici-ency Level, FULL TEST	Below 1st	Count	**331**	220	0	0	0	0	0	0	551
		AR	180,7	47,7	-7,2	-10,6	-13,5	-13,5	-9,9	-5,9	
	1st	Count	122	**1599**	536	0	0	0	0	0	2257
		AR	29,9	**180,3**	26,0	-21,8	-27,7	-27,6	-20,3	-12,0	
	2nd	Count	10	742	**3898**	929	31	0	0	0	5610
		AR	-4,3	42,5	**168,7**	-0,8	-43,7	-44,6	-32,8	-19,4	

Below 1st			**Math Proficiency Level, CLOSED**								**Total**
1st			**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**			
Math Proficiency Level, FULL TEST	3rd	Count	0	18	2132	**8067**	1945	76	6	1	12245
		AR	-9,4	-21,6	37,8	**157,4**	-24,9	-67,4	-50,7	-30,1	
	4th	Count	0	0	34	3942	**12206**	2711	82	7	18982
		AR	-12,4	-29,7	-47,8	16,1	**145,3**	-38,4	-65,0	-39,4	
	5th	Count	0	0	0	44	4763	**12433**	2313	91	19644
		AR	-12,7	-30,4	-49,9	-72,5	-2,1	**145,4**	-14,9	-37,4	
	6th	Count	0	0	0	0	43	3680	**7338**	1185	12246
		AR	-9,4	-22,5	-37,1	-54,6	-68,4	14,9	**151,5**	19,6	
	7th	Count	0	0	0	0	0	20	1787	**3208**	5015
		AR	-5,7	-13,7	-22,5	-33,1	-42,1	-41,3	42,1	**181,1**	
Total		Count	1095	463	2579	6600	12982	18988	18920	11526	4492

Note: Count is Observed Frequencies and AR is Adjusted Standardized Residual

Appendix 3. Comparison of ability levels on the 8th grade between the full test and the closed items only in reading

Below 1st			**Reading Proficiency Level, CLOSED**								**Total**
1st			**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**			
Reading Proficiency Level, FULL TEST	Below 1st	Count	**286**	244	0	0	0	0	0	0	530
		AR	**175,4**	49,1	-8,5	-11,9	-13,6	-12,2	-8,4	-4,5	
	1st	Count	114	**2190**	780	0	0	0	0	0	3084
		AR	25,7	**192,1**	23,6	-29,1	-33,2	-29,8	-20,6	-11,1	
	2nd	Count	3	848	**6557**	1354	8	0	0	0	8770
		AR	-6,5	28,2	**193,3**	-13,4	-58,0	-52,2	-36,1	-19,4	
	3rd	Count	0	1	2224	**11691**	2090	56	2	1	16065
		AR	-10,0	-29,1	8,8	**180,5**	-40,9	-73,3	-51,5	-27,6	
	4th	Count	0	0	8	3858	**14460**	2706	97	12	21141
		AR	-12,0	-34,8	-61,8	-11,1	**165,8**	-36,5	-59,2	-32,5	
	5th	Count	0	0	0	13	4155	**12193**	2338	118	18817
		AR	-11,1	-32,2	-57,4	-80,3	-12,7	**163,8**	3,4	-25,4	
	6th	Count	0	0	0	0	11	2581	**6123**	1132	9847
		AR	-7,5	-21,8	-38,8	-54,5	-61,9	11,6	**166,1**	44,0	
	7th	Count	0	0	0	0	0	1	921	**1710**	2632
		AR	-3,7	-10,7	-19,1	-26,8	-30,6	-27,4	37,7	**169,9**	
Total		Count	1095	403	3283	9569	16916	20724	17537	9481	2973

Note: Count is Observed Frequencies and AR is Adjusted Standardized Residual

Appendix 4. Comparison of ability levels on the 10th grade between the full test and the closed items only in reading

Below 1st			**Reading Proficiency Level, CLOSED**								
1st			**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**			**Total**
Reading Proficiency Level, FULL TEST	Below 1st	Count	**331**	220	0	0	0	0	0	0	551
		AR	**180,7**	47,7	-7,2	-10,6	-13,5	-13,5	-9,9	-5,9	
	1st	Count	122	**1599**	536	0	0	0	0	0	2257
		AR	29,9	**180,3**	26,0	-21,8	-27,7	-27,6	-20,3	-12,0	
	2nd	Count	10	742	**3898**	929	31	0	0	0	5610
		AR	-4,3	42,5	**168,7**	-0,8	-43,7	-44,6	-32,8	-19,4	
	3rd	Count	0	18	2132	**8067**	1945	76	6	1	12245
		AR	-9,4	-21,6	37,8	**157,4**	-24,9	-67,4	-50,7	-30,1	
	4th	Count	0	0	34	3942	**12206**	2711	82	7	18982
		AR	-12,4	-29,7	-47,8	16,1	**145,3**	-38,4	-65,0	-39,4	
	5th	Count	0	0	0	44	4763	**12433**	2313	91	19644
		AR	-12,7	-30,4	-49,9	-72,5	-2,1	**145,4**	-14,9	-37,4	
	6th	Count	0	0	0	0	43	3680	**7338**	1185	12246
		AR	-9,4	-22,5	-37,1	-54,6	-68,4	14,9	**151,5**	19,6	
	7th	Count	0	0	0	0	0	20	1787	**3208**	5015
		AR	-5,7	-13,7	-22,5	-33,1	-42,1	-41,3	42,1	**181,1**	
Total		Count	1095	463	2579	6600	12982	18988	18920	11526	4492

Note: Count is Observed Frequencies and AR is Adjusted Standardized Residual

ATOM'S STRUCTURE – EMPLOYEE AND EMPLOYER FEEDBACK, SURVEY SITE

Máté Pusker
CIVIL Plc
mate.pusker@gmail.com

Bence Gergely
ELTE Eötvös Loránd University, Doctoral School of Psychology
Károli University of the Reformed Church in Hungary
gergely.bence98@outlook.com

Szabolcs Takács
Károli University of the Reformed Church in Hungary
takacs.szabolcs.dr@gmail.com

Summary

Background and Aims: Presenting the fundamentals of ATOM functionalities to give the readers insight into how the three types of users (namely employees, employers, and analyst experts) might work with ATOM in their respective practice.
Methods: Showing selected main screenshots and interpreting their related functionalities in terms of automated manpower selection.
Discussion and Conclusions: It is concluded that all the necessary sets of employee, employer, and expert functions can be adequately accessed in the software to support its users and assist them in the recruitment process.
Keywords: recruitment, selection campaigns, automatized workforce selection, personalized feedback

DOI: 10.17627/ALKPSZICH.2023.3.55

Introduction

ATOM's main goal and advantage are that it draws upon today's technological prospects to host different key steps of the recruitment process on one platform while using a unique methodology in the recruitment cycle. Utilizing information technology throughout the recruitment process has been broadly adopted (Nikolaou, 2014; McCarthy et al., 2017), which leads us to consider the next step in technological evolution, namely using machine learning, automation, and artificial intelligence in recruitment.

While constructing ATOM, it was very important to create an application that addresses goals, requirements, and trends that immensely affect today's recruitment. One of these aspects is employer branding (Nikolaou, 2014; McCarthy et al., 2017) since the potential workforce gathers information regarding a company and HR functionality via recruitment techniques (Woods, 2020; Nikolaou, 2021). As a result, optimizing such processes is a fundamentally important guideline for employers and soon-to-be employees.

Selecting potential personnel has been at the centre of attention since the foundation of Applied Psychology (Polyhart, et al.,2017). Considering the expansion of requirements and conditions when applying for a position, it has become critical to understand what sort of technology and methodology should be used to effectively measure applicants' knowledge, abilities, and other characteristics (Potočnik et al., 2021).

Detailed job profiles are the basis upon which a set of requirements can be defined that collaborate to measure applicants' competencies. In addition, different tasks such as logical, reading comprehension, and situational assignment have been digitalized, accelerating aptitude testing (Tippins, 2015); moreover, gathering data with the use of the internet has become easily accessible (Gosling, 2004).

Involving self-reported personality questionnaires, supporting the recruitment process, has also become popular (Ryan et al., 2015) and represents extra means to measure applicants. Although using such questionnaires is viewed and supported differently among professionals (Diekmann & König, 2015; Risavy et al., 2019), every additional information in connection with applicants assists recruitment professionals in selecting suitable employees for open positions (Phillips & Gully, 2015).

On the one hand, it is in the employer's best interest to attract as many outstanding candidates as possible (Collins & Kanar, 2014). On the other hand, it is presumably an advantage if applicants can expect a fair selection process when applying for a role. Bad candidate experience might have a negative impact on the employer's brand that quickly leads to a disadvantage, resulting in a reduced number of quality applications, and all this because of a malfunctioning recruitment process (Miles & McCamey, 2018). One of the most important aspects of well-designed recruitment management systems (RMS) is adequate information and feedback toward candidates (McCarthy et al., 2018; Rozario et al., 2019). Jobseekers may invest a lot of time and energy in following through with the application process and providing information about their person-job fit, so it is essential to provide them feedback.

Recruitment professionals evaluate candidates based on previously defined criteria with the support of even maybe industry-specific professionals if the role requires it. This evaluation process is time-consuming that prolongs the recruitment process itself. Moreover, objective evaluation is challenging since all CVs differ in layout, structure, and professional content (Faliagka et al., 2012).

When selecting a suitable candidate for a position, it is difficult to determine if we can accurately predict the right person-job fit. Since objectivity and accuracy of test evaluation show a substantial difference between machine learning and people (Youyou et al., 2015), while creating the ATOM framework, it was of great importance to embed a module that can successfully make an objective prediction (Gergely & Takács, this special issue, *Methods*).

ATOM has machine learning embedded in its core framework that enables additional functions to be housed within the application. Such extensions are automated feedback for employers and employees as their test scores are automatically evaluated and summarized in reports. This supports the basic requirement for candidates to receive personalized, dynamic feedback on their test performance (Tippins, 2015).

ATOM's goal is:

- to accelerate the recruitment process;
- to provide HR professionals with objective information and prediction regarding person-job fit;
- to provide automated, personalized, adequate feedback to both employers and employees.

The fundamentals of the core ATOM modularity are accessible in our thematic articles (Gergely & Takács, this special issue, *Methods*).

Key functionalities of ATOM

In the following, the key functionalities of ATOM are presented, which provide different users with interaction possibilities while progressing in the recruitment process. Introduction to these functionalities occurs chronologically, explaining the methodology behind this recruitment management system. For each set of functionalities corresponding to the four primary windows of ATOM, visual illustrations (screenshots) are presented below.

The four primary windows

The opening screen contains the following four primary windows by the help of which the user can have access to all the functionalities that ATOM provides (as all the texts within ATOM are in Hungarian, the Hungarian names are also indicated in brackets):

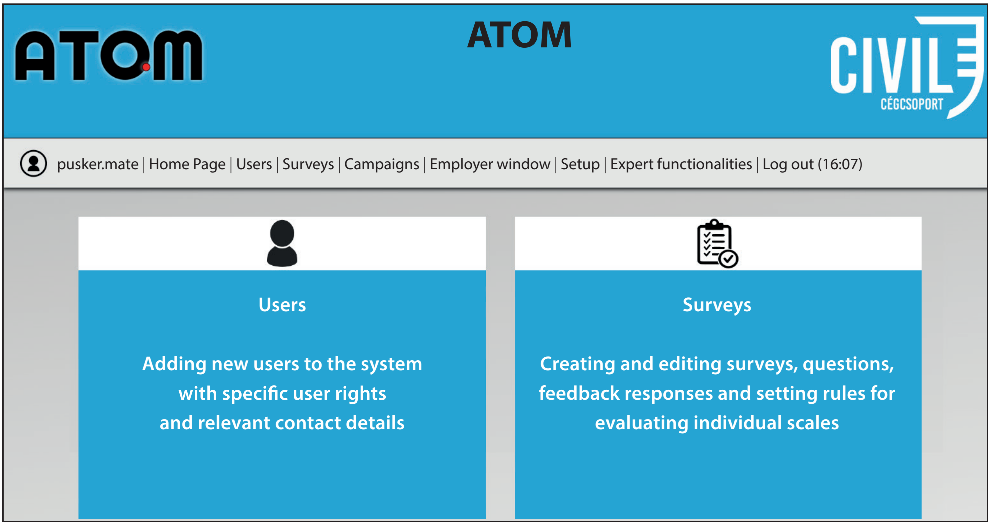
Users (Felhasználók): provides possibilities to import new users to have access to certain functions depending on the various credentials we provide them.

Surveys (Kérdőívek): provides access allowing users to manage and add new tests (or other instruments) to the system.

Setup (Beállítások): provides possibilities to set up the server and for inserting external data to be evaluated by the core of ATOM. These loading and running functions belong to the expert's functionalities.

Campaigns (Kampányok): It provides possibilities to tailor recruitment campaigns and add tests to them to effectively screen candidates for a specific position based on the client's requirements.

To have a detailed presentation, we have all available functionalities in the menu on the opening screen. We can see that the functionalities presented in the four primary windows can also be found in the menu located at the top of the opening screen. This menu also lists the Employer window *(Munkáltatói felület)* for its set of functionalities and the Expert functionalities *(Szakértői felület).* All functionalities are available in the four primary windows and in the menu of the opening screen are designed to be managed by users such as employers and analyst experts.


ATOM
ATOM
CIVIL
CÉGCSOPORT
pusker.mate | Home Page | Users | Surveys | Campaigns | Employer window | Setup | Expert functionalities | Log out (16:07)

Users

Adding new users to the system with specific user rights and relevant contact details

Surveys

Creating and editing surveys, questions, feedback responses and setting rules for evaluating individual scales

Figure 1. The opening screen of ATOM with the four primary windows[1]

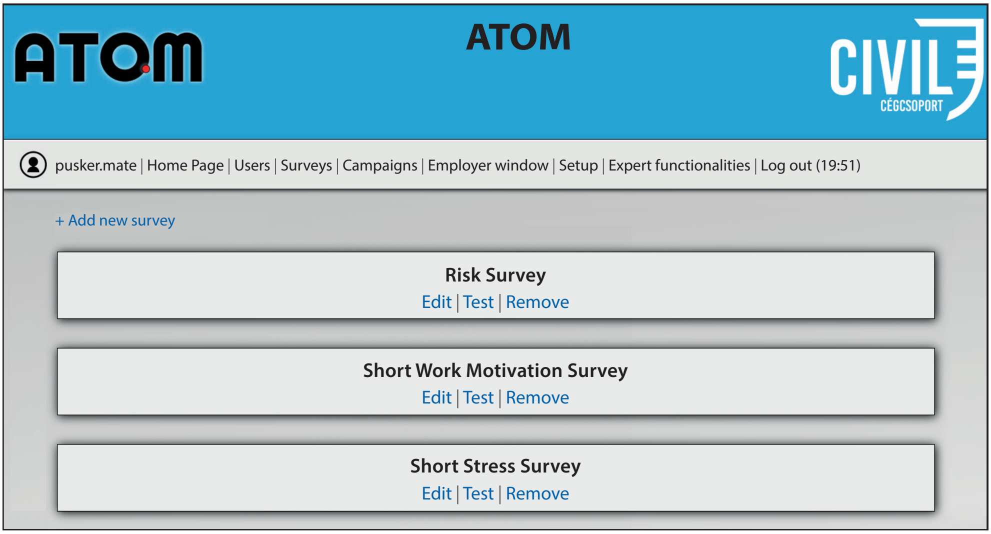
The main dialog boxes

Surveys (Kérdőívek)

This set of functions enables us to further access or expand on available tests, which can later be utilized to screen candidates when applying for a role. When selecting this option, we receive a pool of existing questionnaires ready to be included in a new recruitment campaign. Adding a new test will further expand the available tools that offer various scopes for screening candidates.

There are cloud-based Survey applications (Google Forms, Onlinekérdőív.hu) available, however, ATOM offers two new key aspects to surveying candidates.

[1] There is currently no English version of the ATOM service, *Figures 1–11* published in the present article are illustrations, i.e. visual designs based on the Hungarian original. They are published in order to illustrate the functioning of the ATOM system and to help the English-speaking reader to better understand its details. For the original Hungarian version of these screenshots, please refer to *Appendices 1–11.*


ATOM
ATOM
CIVIL
CÉGCSOPORT
pusker.mate | Home Page | Users | Surveys | Campaigns | Employer window | Setup | Expert functionalities | Log out (19:51)
+ Add new survey

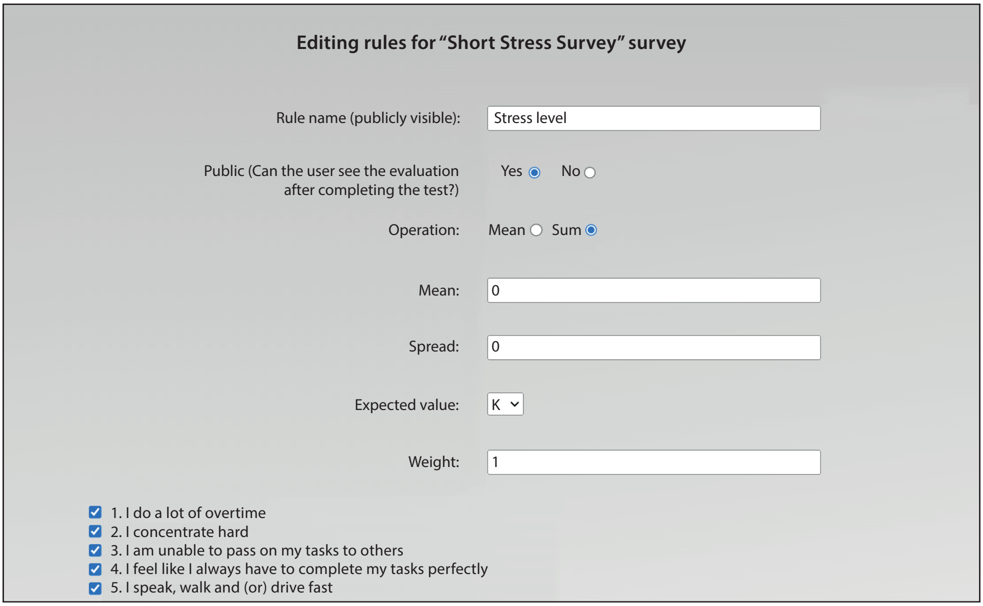
Risk Survey
Edit | Test | Remove

Short Work Motivation Survey
Edit | Test | Remove

Short Stress Survey
Edit | Test | Remove

Figure 2. Secondary window for accessing surveys

Firstly, each question can be instantly assigned to the adequate scale they belong to. These scales are the building blocks for defining and measuring values, which are essentially the indicators for the person-job fit.



Editing rules for "Short Stress Survey" survey

Rule name (publicly visible): Stress level

Public (Can the user see the evaluation after completing the test?) Yes ● No ○

Operation: Mean ○ Sum ●

Mean: 0

Spread: 0

Expected value: K

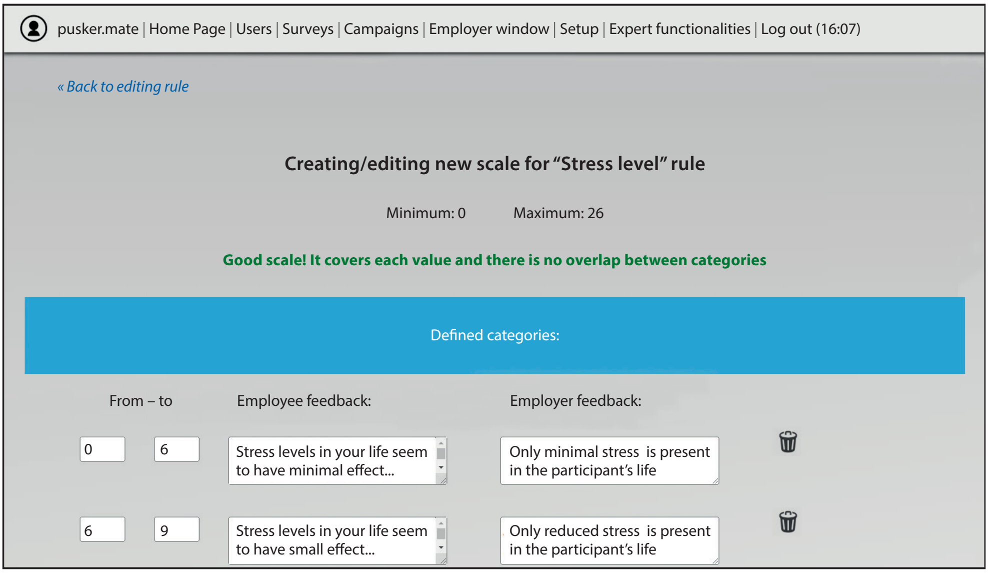
Weight: 1

☑ 1. I do a lot of overtime
☑ 2. I concentrate hard
☑ 3. I am unable to pass on my tasks to others
☑ 4. I feel like I always have to complete my tasks perfectly
☑ 5. I speak, walk and (or) drive fast

Figure 3. Assigning questions to certain scales

Secondly, it is key to have adequate feedback, which can be communicated to each person who undertakes the screening process. Communicating results with survey participants can easily be optimized and made fast. A real-life application can result in three distinct categories based on low/middle/high values with individual evaluations attached to them, providing personalized and automated feedback to participants based on their answers.

While inserting employee feedback into specific scales and dimensions, we integrated a section for employers. So, after an applicant has completed a survey, they automatically receive personal feedback. At the same time, a report is also generated automatically in the system so employers can see the candidate's results. The only difference between the two types of feedback is that the employers receive a more detailed summary of the particular scale and its results.


pusker.mate | Home Page | Users | Surveys | Campaigns | Employer window | Setup | Expert functionalities | Log out (16:07)
« Back to editing rule

Creating/editing new scale for "Stress level" rule

Minimum: 0 Maximum: 26

Good scale! It covers each value and there is no overlap between categories

Defined categories:

From – to		Employee feedback:	Employer feedback:
0	6	Stress levels in your life seem to have minimal effect...	Only minimal stress is present in the participant's life
6	9	Stress levels in your life seem to have small effect...	Only reduced stress is present in the participant's life

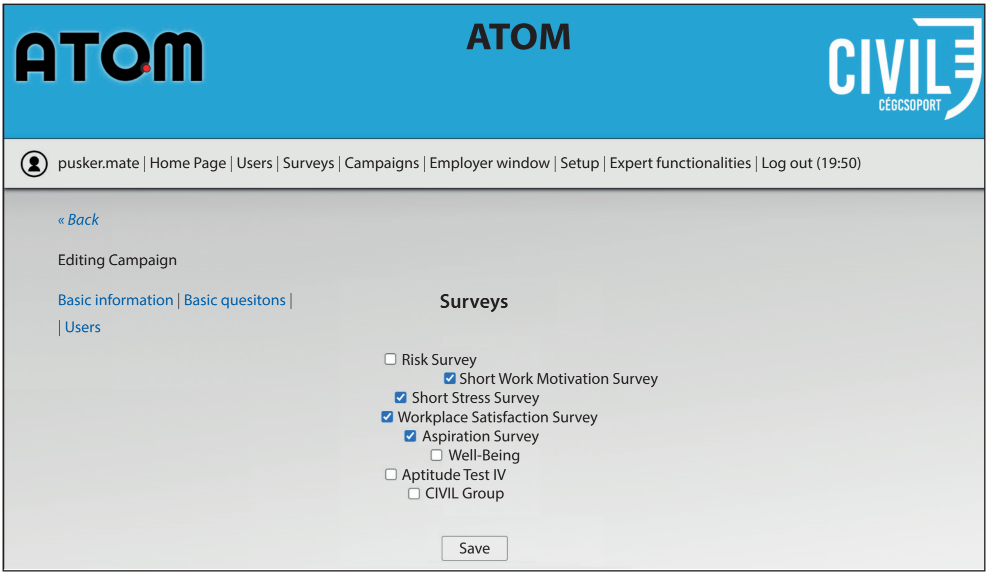
Figure 4. Automated feedback for both employees and employers

Campaigns (Kampányok)

When a client wants to recruit for a specific position, this secondary window allows us to create recruitment campaigns for them. This brand-new campaign will be assigned with an automatically generated link, which can be easily accessed online by all candidates in the active recruitment period, who visit the job description for the open position. The previously introduced questionnaires are the backbone of these campaigns. After professional revision, adequate tests can be easily assigned to a recruitment campaign.

Since certain positions require complex competencies, we need to obtain as much information about candidates as possible. Hence, several different testing instruments can be included in a recruitment campaign. This will allow employers to understand the applicants better to make an objective, data-driven decision that focuses on previously defined job criteria.

After selecting a variety of tests that we consider adequate to measure person-job fit for a specific position, we can further expand on gathering valuable information about the quality of our candidate pool and their expectations. This will also be part of the complex criteria system based on which we can evaluate person-job fit more thoroughly. Such criteria can be, for instance, possessing a driving license, foreign language knowledge, computer skills, or other relevant certificates. We gather all this information and consider them to be part of so-called basic criteria since it can effectively direct the attention and resources of recruitment professionals, especially when experiencing a high volume of candidates applying for positions, supporting their objective decision-making process.


ATOM
ATOM
CIVIL
CÉGCSOPORT
pusker.mate | Home Page | Users | Surveys | Campaigns | Employer window | Setup | Expert functionalities | Log out (19:50)
« Back
Basic information | Basic quesitons |
| Users
Save

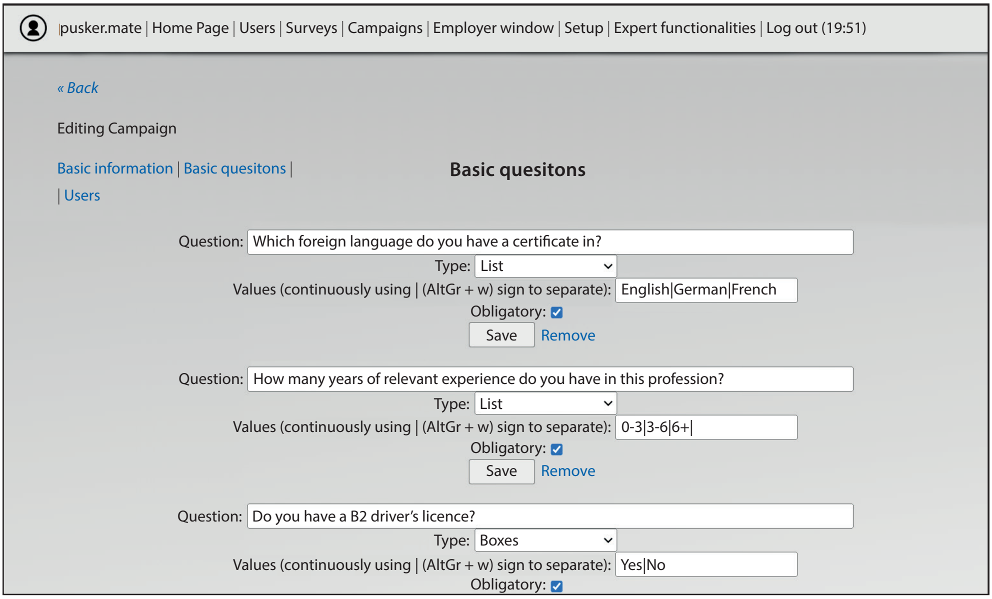
Editing Campaign

Surveys

☐ Risk Survey
☑ Short Work Motivation Survey
☑ Short Stress Survey
☑ Workplace Satisfaction Survey
☑ Aspiration Survey
☐ Well-Being
☐ Aptitude Test IV
☐ CIVIL Group

Figure 4. Automated feedback for both employees and employers

It is highly important to gain insight into the professional background of candidates. Information about relevant studies, prior employment, and experiences are all found in CVs. Collecting these are also part of the embedded recruitment process in ATOM. This will be explained in detail in the upcoming section when we introduce the Employee window.


pusker.mate | Home Page | Users | Surveys | Campaigns | Employer window | Setup | Expert functionalities | Log out (19:51)
« Back
Basic information | Basic quesitons |
| Users
Save
Remove
Save
Remove

Editing Campaign

Basic quesitons

Question: Which foreign language do you have a certificate in?
Type: List
Values (continuously using | (AltGr + w) sign to separate): English|German|French
Obligatory: ☑

Question: How many years of relevant experience do you have in this profession?
Type: List
Values (continuously using | (AltGr + w) sign to separate): 0-3|3-6|6+|
Obligatory: ☑

Question: Do you have a B2 driver's licence?
Type: Boxes
Values (continuously using | (AltGr + w) sign to separate): Yes|No
Obligatory: ☑

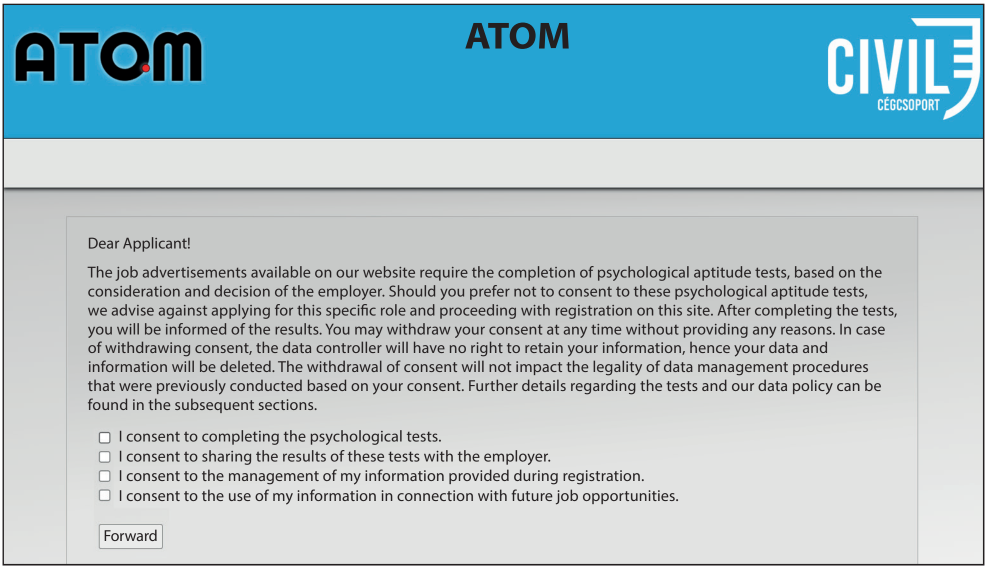
Figure 6. Basic criteria can be integrated based on client requirements

Our selected questionnaires support our aim of understanding our candidates and measuring the quality of person-job fit. After gathering all additional questions to provide a primary criterion for choosing the suitable person for the job, the campaign is ready. Once a candidate progresses through the basic criteria questions, uploads necessary documents (CV, certificates), and fills in the questionnaires, the application for the opening is registered and saved in the database.

Employees' functionality (Munkavállalói felület)

As mentioned before, all campaigns have a designated link that can be published on all platforms available for the client to recruit new workforce. It is important to note that we called this platform "Employee function" on purpose. These "campaigns" can also focus on organizational surveys such as employee satisfaction, organizational commitment, etc. Once opening the link, each candidate will be directed to ATOM and, more precisely, to the Employee window's opening screen. As introduced before, the aim of guiding candidates to this window is to unify and simplify the application process by merging several recruitment steps (gathering CVs, first pre-screening call, testing phase).

Once clicking on the available link, the candidate will be directed to the opening screen. First, registration will be necessary, and the account will save all required information. This makes the application process more manageable since it enables candidates to revisit their account and their progress in applying for a job. This step cannot be completed without consenting to relevant GDPR protocols and guidelines for protecting applicants.



Dear Applicant!

The job advertisements available on our website require the completion of psychological aptitude tests, based on the consideration and decision of the employer. Should you prefer not to consent to these psychological aptitude tests, we advise against applying for this specific role and proceeding with registration on this site. After completing the tests, you will be informed of the results. You may withdraw your consent at any time without providing any reasons. In case of withdrawing consent, the data controller will have no right to retain your information, hence your data and information will be deleted. The withdrawal of consent will not impact the legality of data management procedures that were previously conducted based on your consent. Further details regarding the tests and our data policy can be found in the subsequent sections.

□ I consent to completing the psychological tests.
□ I consent to sharing the results of these tests with the employer.
□ I consent to the management of my information provided during registration.
□ I consent to the use of my information in connection with future job opportunities.

Figure 7. GDPR compliance

The application process is relatively straightforward:

1. Registration on ATOM's Employee window
2. Using the credentials provided in the registration to enter the candidate account
3. Providing information to basic questions relevant to the role and the company
4. Uploading necessary documents
5. Completing the survey to finish the application process

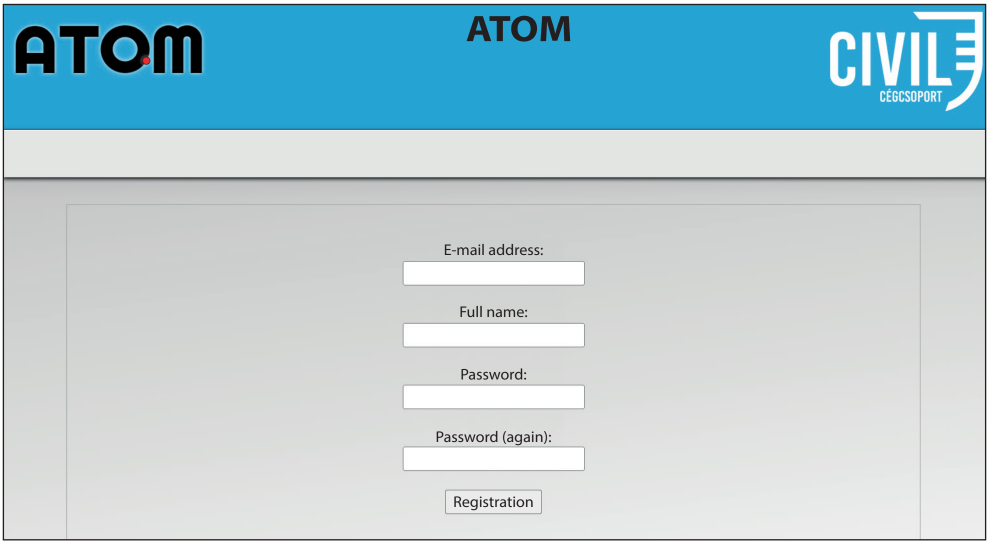
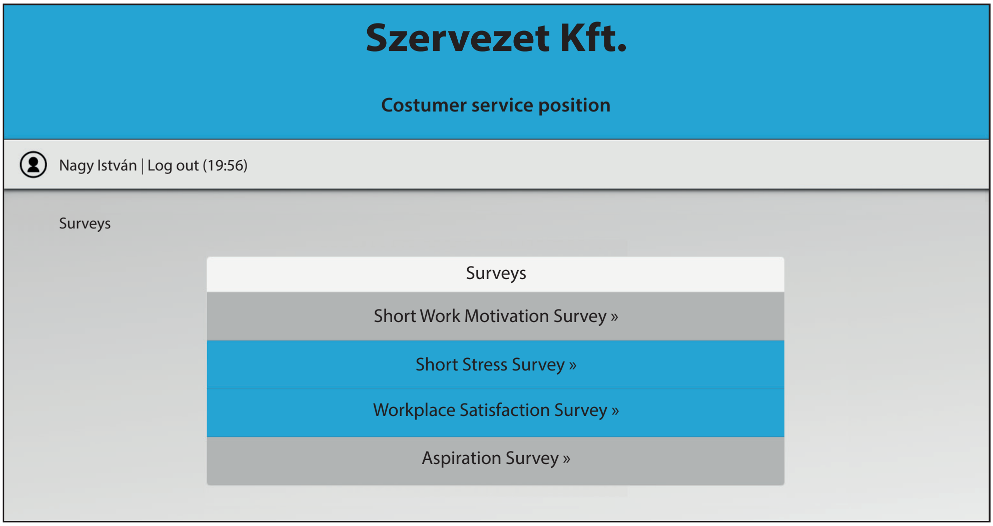


Figure 8. Registration screen. Anonymous participation is included for organizational surveys

As mentioned before, each role may have different testing protocols included in the application process that depends on HR professionals and their professional insights regarding the position. Each test can be answered independently. When more tests are required for a position, it is easy to revisit one's account and continue the next test to complete the application process. However, it is important to note that each test can be answered only once. When logging in to an account, all finished tests appear grey as they have already been answered. However, applicants will still have access to the personal feedback they automatically receive after finishing a test. Once finishing the assigned tests, applicants will automatically receive personal feedback based on their survey answers.



Szervezet Kft.

Costumer service position

Nagy István | Log out (19:56)

Surveys

Surveys

Short Work Motivation Survey »

Short Stress Survey »

Workplace Satisfaction Survey »

Aspiration Survey »

Figure 9. Finished (grey) and unfinished (blue) tests in the Employee window

It is important to note that these feedbacks are not from evaluations on their performance. These are strictly constructive comments, which we can provide based on the particular dimensions of a certain test. These generally fall into *"below average"*, *"average"*, and *"above average"* categories.

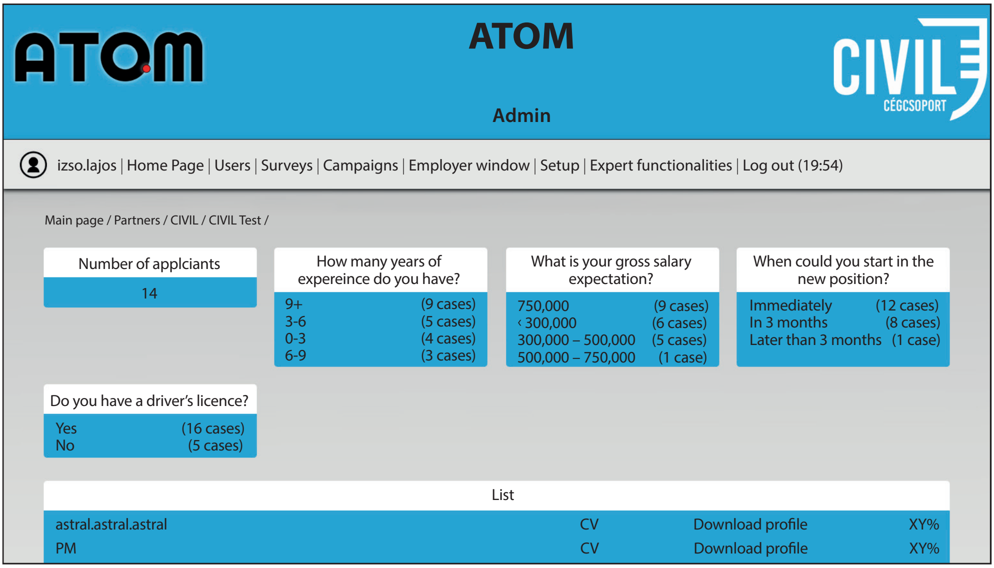
Employers' functionality (Munkáltatói felület)

The purpose of this function is to collect and store applicant information according to campaigns in a simplified and organized structure. Building on the unique methodology of the previously detailed recruitment process, we obtain a set of predetermined criteria that provide input to the core of ATOM that results in predicting the person-job fit for each applicant. This rank order makes ATOM and the Employer function unique and provides its users with considerable benefits.

It is available to sort through applicants based on different conditions for the constantly changing labor market. Should a specific campaign attract a high volume of competent applicants – an ideal scenario –it is easy to organize candidates, starting the list with the most competent ones. It becomes equally important in non-ideal market conditions when a lack of a professional workforce results in recruitment challenges, which applicants must avoid when striving to fill open positions.

The Employer function can be described in two main components based on the initial structure of the Campaigns. The content on this secondary window starts with a general description of a particular campaign and lists a summary based on the criteria defined by the client. These can be, for example:

- Number of applicants for the position
- Relevant experience related to the position
- Notice period
- Salary expectation
- Etc.


ATOM
ATOM
Admin
CIVIL
CÉGCSOPORT
izso.lajos | Home Page | Users | Surveys | Campaigns | Employer window | Setup | Expert functionalities | Log out (19:54)
Main page / Partners / CIVIL / CIVIL Test /

Number of applciants
14

How many years of expereince do you have?	
9+	(9 cases)
3-6	(5 cases)
0-3	(4 cases)
6-9	(3 cases)

What is your gross salary expectation?	
750,000	(9 cases)
‹ 300,000	(6 cases)
300,000 – 500,000	(5 cases)
500,000 – 750,000	(1 case)

When could you start in the new position?	
Immediately	(12 cases)
In 3 months	(8 cases)
Later than 3 months	(1 case)

Do you have a driver's licence?	
Yes	(16 cases)
No	(5 cases)

List			
astral.astral.astral	CV	Download profile	XY%
PM	CV	Download profile	XY%

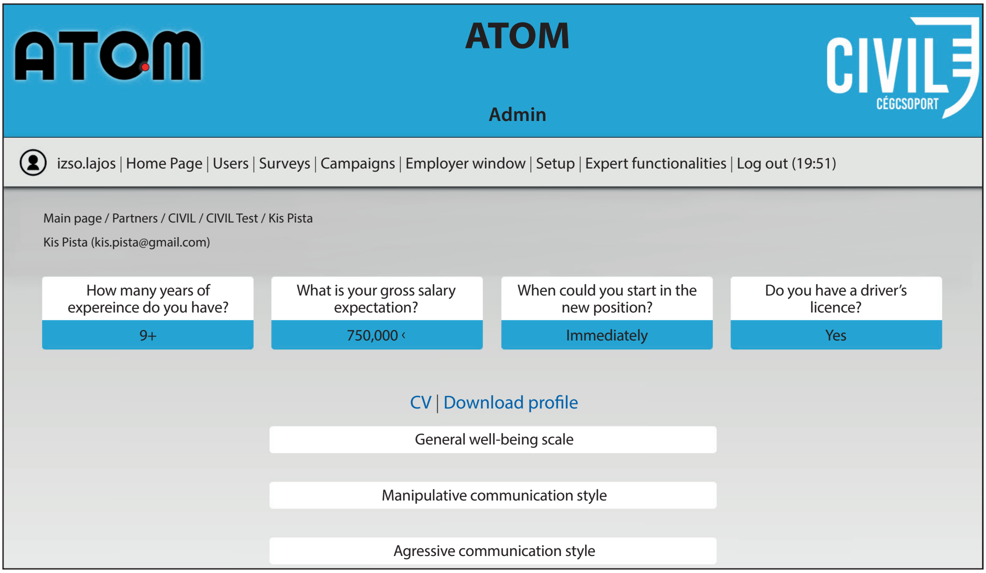
Figure 10. Pinned summarizations of basic criteria

These summaries build on the requirements and key aspects of a role provided by HR professionals that understand the position at hand.

After this summary, we find the rank order of candidates, which is the key element for this window. The necessary input for making predictions regarding person-job fit derives from the testing sequence built in each campaign. This person-job fit indicator is automatically calculated in ATOM's core (Gergely & Takács, this special issue, *Concurrent algorithms, hyperparameters, and cross-validation*) and is listed for each applicant.

This rank list also pertains information about each candidate, such as their CV and profile. The profile contains the employer feedback explained previously, generated based on applicants' answers in the testing phase. Clicking on a person's name on this list will provide us with their answers for the base criteria questions and the results for each dimension in a survey assigned to test the candidate.

Having opened a candidate profile, we can download their uploaded CV. We can also download the profile seen on the page that lists the relevant feedback assigned to their score (*"below average"*, *"average"*, and *"above average"*) in connection with each dimension. As explained earlier, this feedback provides a more in-depth summary and is more direct for professionals to evaluate and integrate into their objective decision-making process.



ATOM

Admin

izso.lajos | Home Page | Users | Surveys | Campaigns | Employer window | Setup | Expert functionalities | Log out (19:51)

Main page / Partners / CIVIL / CIVIL Test / Kis Pista

Kis Pista (kis.pista@gmail.com)

How many years of expereince do you have?
9+

What is your gross salary expectation?
750,000 ‹

When could you start in the new position?
Immediately

Do you have a driver's licence?
Yes

CV | Download profile

General well-being scale

Manipulative communication style

Agressive communication style

Figure 11. Available online candidate profile

Experts' functionality (Szakértői felület)

An independent Experts' functionality (Szakértői felület) has been integrated into the system, enabling professionals to analyze complex databases from external sources and utilize ATOM's core algorithms to evaluate data.

Discussion

ATOM's core and its properties can significantly affect how we conduct recruitment processes. It aims to simplify and improve the process while featuring a new approach to screening candidates. Available functionalities for candidates make it easy to follow a user journey while receiving automatic and personalized feedback for investing time and effort to complete the process. This serves as a unique selling point that benefits the employer's brand. Employers gain easy access to applicant information while obtaining additional test results about applicants. This allows HR and recruitment professionals to understand candidates better, resulting in a more unbiased and objective evaluation. Promoting this, predictions on person-job fit quality are also available to support professionals. Generating reports and customizing feedback are standard requirements that still demand precious time from experts hindering them in their practice. ATOM aims to serve as a support system that benefits experts, employers, employees, and applicants.

Összefoglalás

Az ATOM szerkezete – visszajelzés és kérdőívek a munkavállaló és a munkáltató részére

Háttér és célkitűzések: Az ATOM-funkciók alapvetéseinek bemutatása, hogy az olvasók betekintést nyerjenek abba, a háromféle felhasználó (nevezetesen a munkavállaló, a munkáltató és az elemző-szakértő) miként dolgozhat az ATOM-mal a saját munkaterületén.
Módszer: A kiválasztott és legfontosabb képernyőképek bemutatása, valamint a hozzájuk kapcsolódó funkciók értelmezése az automatizált munkaerő-kiválasztás szempontjából.
Megbeszélés és következtetések: Arra a következtetésre jutottunk, hogy a munkavállalói, munkáltatói és szakértői funkciók szükséges készlete elérhető a szoftverben, hogy támogassa a felhasználókat, és segítse őket a munkaerő-felvételi folyamatban.
Kulcsszavak: toborzás, kiválasztási kampányok, automatizált munkaerő-kiválasztás, személyre szabott visszajelzés

References of this Special Issue

Gergely, B., & Takács, Sz. (2023). ATOM – a flexible multi-method machine learning framework for predicting occupational success. *Alkalmazott Pszichológia*, *25*(3), 15–30.

References

Collins, C. J., & Kanar, A. M. (2014). Employer brand equity and recruitment research. In K. Y. T. Yu, & D. M. Cable (Eds.), *The Oxford Handbook of Recruitment* (pp. 284–297). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199756094.013.0016

Diekmann, J., & König, C. (2015). Personality testing in personnel selection: Love it? Leave it? Understand it! In I. Nikolaou, & J. K. Oostrom (Eds.), *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice* (pp. 117–135). Psychology Press.

Faliagka, E., Tsakalidis, A., & Tzimas, G. (2012). An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet Research*, *22*(5), 551–568. https://doi.org/10.1108/10662241211271545

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, *59*(2), 93–104. https://doi.org/10.1037/0003-066X.59.2.93

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., & Ahmed, S. M. (2017). Applicant perspectives during selection: a review addressing "So What?," "What's New?," and "Where to Next?" *Journal of Management*, *43*(6), 1693–1725. https://doi.org/10.1177/0149206316681846

Miles, S. J., & McCamey, R. (2018). *The candidate experience: Is it damaging your employer brand?* Business Horizons. https://doi.org/10.1016/j.bushor.2018.05.007

Nikolaou, I. (2014). Social networking web sites in job search and employee recruitment. *International Journal of Selection and Assessment, 22*(2), 179–189. https://doi.org/10.1111/ijsa.12067

Phillips, J. M., & Gully, S. M. (2015). Multilevel and strategic recruiting: Where have we been, where can we go from here? *Journal of Management, 41*(5), 1416–1445. https://doi.org/10.1177/0149206315582248

Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the Supreme Problem: 100 years of selection and recruitment at the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 291–304. https://doi.org/10.1037/apl0000081

Potočnik, K., Anderson, N. R., Born, M., Kleinmann, M., & Nikolaou, I. (2021). Paving the way for research in recruitment and selection: Recent developments, challenges, and future opportunities. *European Journal of Work and Organizational Psychology, 30*(2), 159–174. https://doi.org/10.1080/1359432X.2021.1904898

Risavy, S. D.; Fisher, P. A.; Robie, C., & König, C. J. (2019). Selection Tool Use: A Focus on Personality Testing in Canada, the United States, and Germany. *Personnel Assessment and Decisions*, *5*(1), Article 4. https://scholarworks.bgsu.edu/pad/vol5/iss1/4 https://doi.org/10.25035/pad.2019.01.004

Rozario, S. D., Venkatraman, S., & Abbas, A. (2019). Challenges in Recruitment and Selection Process: An Empirical Study. *Challenges*, *10*(2), 35. http://dx.doi.org/10.3390/challe10020035

Ryan, A. M., Inceoglu, I., Bartram, D., Golubovich, J., Grand, J.A., Reeder, M., Derous, E., Nikolaou, I., & Yao, X. (2015). Trends in testing: Highlights of a global survey. In I. Nikolaou, & J. Oostrom (Eds.), *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice* (pp. 136–153). Psychology Press. https://doi.org/10.4324/9781315742175

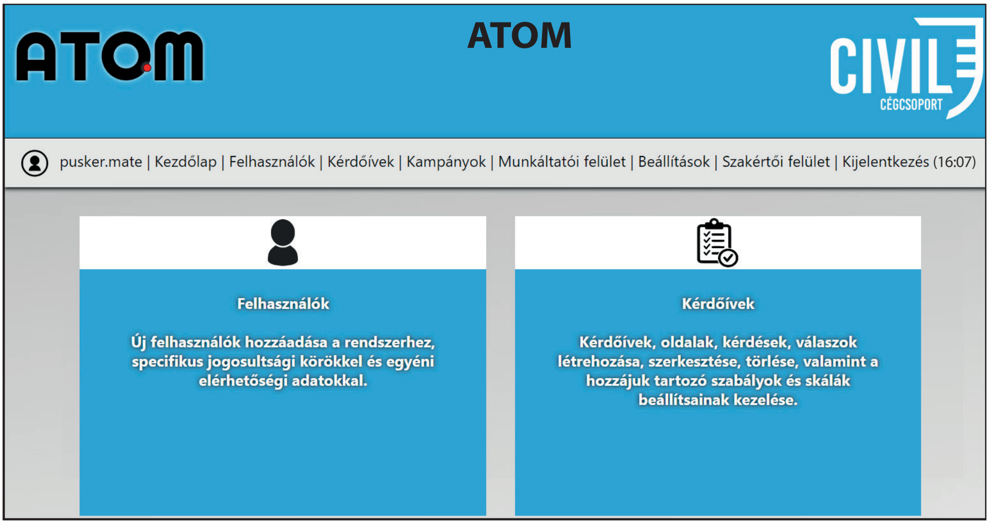
Tippins, N. T. (2015). Technology and assessment in selection. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 551–582. https://doi.org/10.1146/annurev-orgpsych-031413-091317

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, *29*(1), 64–77. https://doi.org/10.1080/1359432X.2019.1681401

Appendices

Appendix 1.


ATOM
ATOM
CIVIL CÉGCSOPORT
pusker.mate | Kezdőlap | Felhasználók | Kérdőívek | Kampányok | Munkáltatói felület | Beállítások | Szakértői felület | Kijelentkezés (16:07)

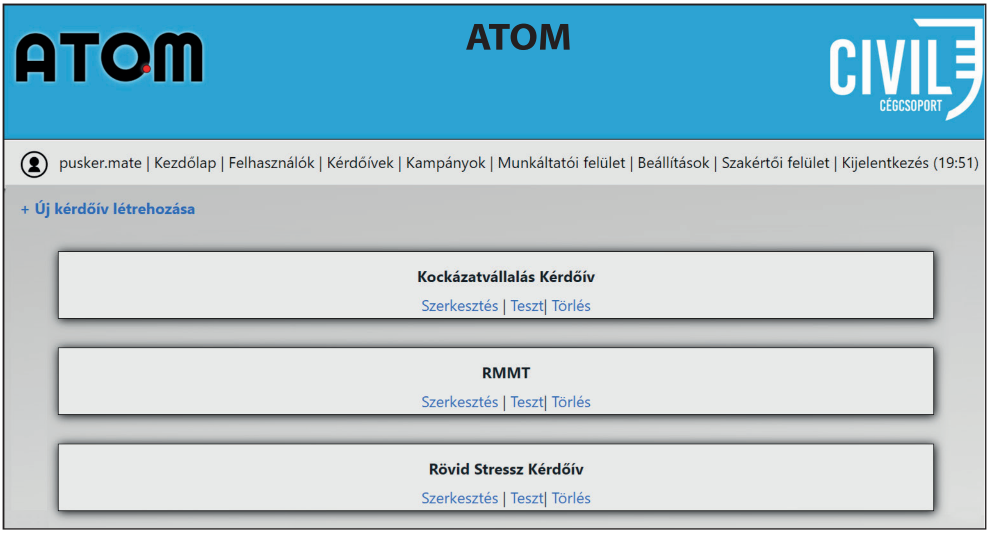
Felhasználók

Új felhasználók hozzáadása a rendszerhez, specifikus jogosultsági körökkel és egyéni elérhetőségi adatokkal.

Kérdőívek

Kérdőívek, oldalak, kérdések, válaszok létrehozása, szerkesztése, törlése, valamint a hozzájuk tartozó szabályok és skálák beállítsainak kezelése.

Appendix 2.


ATOM
ATOM
CIVIL CÉGCSOPORT
pusker.mate | Kezdőlap | Felhasználók | Kérdőívek | Kampányok | Munkáltatói felület | Beállítások | Szakértői felület | Kijelentkezés (19:51)
+ Új kérdőív létrehozása

Kockázatvállalás Kérdőív

Szerkesztés | Teszt| Törlés

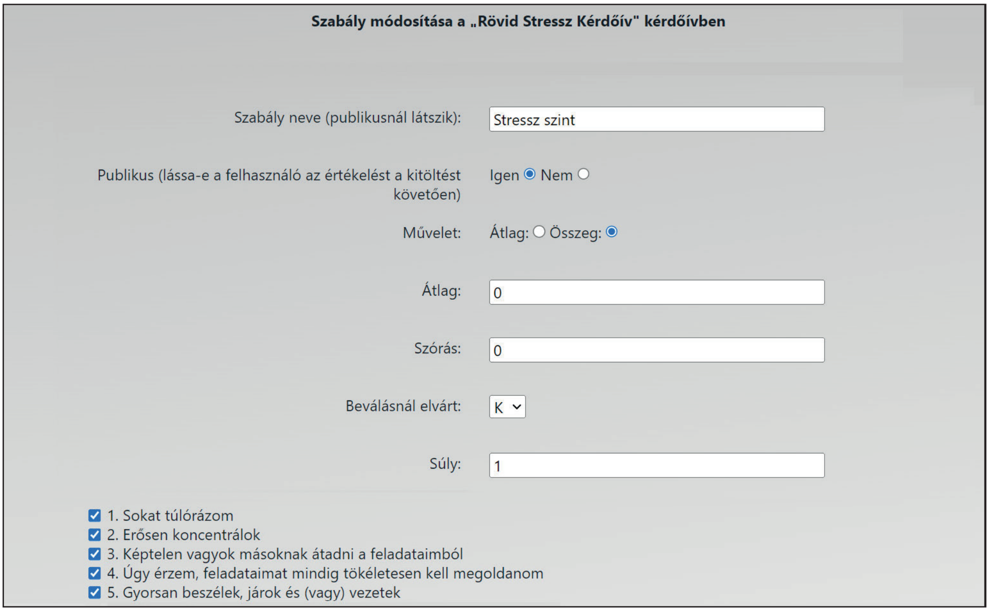
RMMT

Szerkesztés | Teszt| Törlés

Rövid Stressz Kérdőív

Szerkesztés | Teszt| Törlés

Appendix 3.



Szabály módosítása a „Rövid Stressz Kérdőív" kérdőívben

Szabály neve (publikusnál látszik): Stressz szint

Publikus (lássa-e a felhasználó az értékelést a kitöltést követően) Igen ● Nem ○

Művelet: Átlag: ○ Összeg: ●

Átlag: 0

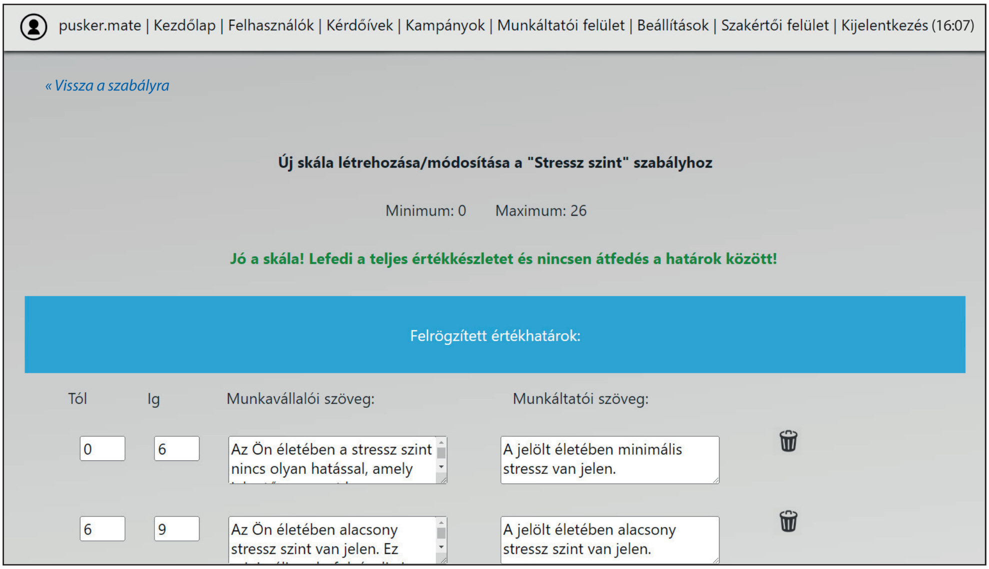
Szórás: 0

Beválásnál elvárt: K

Súly: 1

☑ 1. Sokat túlórázom
☑ 2. Erősen koncentrálok
☑ 3. Képtelen vagyok másoknak átadni a feladataimból
☑ 4. Úgy érzem, feladataimat mindig tökéletesen kell megoldanom
☑ 5. Gyorsan beszélek, járok és (vagy) vezetek

Appendix 4.



pusker.mate | Kezdőlap | Felhasználók | Kérdőívek | Kampányok | Munkáltatói felület | Beállítások | Szakértői felület | Kijelentkezés (16:07)
« Vissza a szabályra

Új skála létrehozása/módosítása a "Stressz szint" szabályhoz

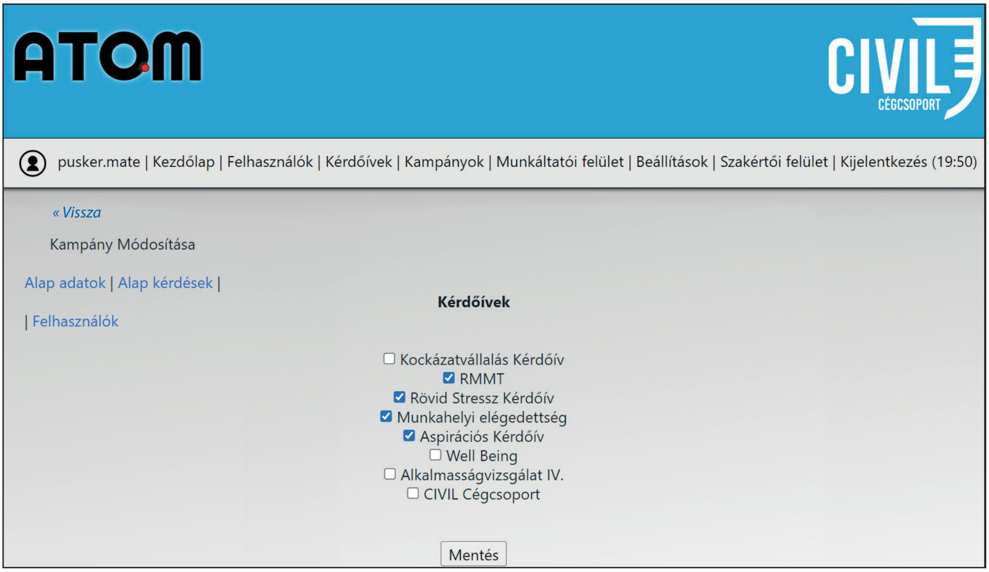
Minimum: 0 Maximum: 26

Jó a skála! Lefedi a teljes értékkészletet és nincsen átfedés a határok között!

Felrögzített értékhatárok:

Tól	Ig	Munkavállalói szöveg:	Munkáltatói szöveg:
0	6	Az Ön életében a stressz szint nincs olyan hatással, amely	A jelölt életében minimális stressz van jelen.
6	9	Az Ön életében alacsony stressz szint van jelen. Ez	A jelölt életében alacsony stressz szint van jelen.

Appendix 5.



pusker.mate | Kezdőlap | Felhasználók | Kérdőívek | Kampányok | Munkáltatói felület | Beállítások | Szakértői felület | Kijelentkezés (19:50)

« Vissza

Kampány Módosítása

Alap adatok | Alap kérdések |

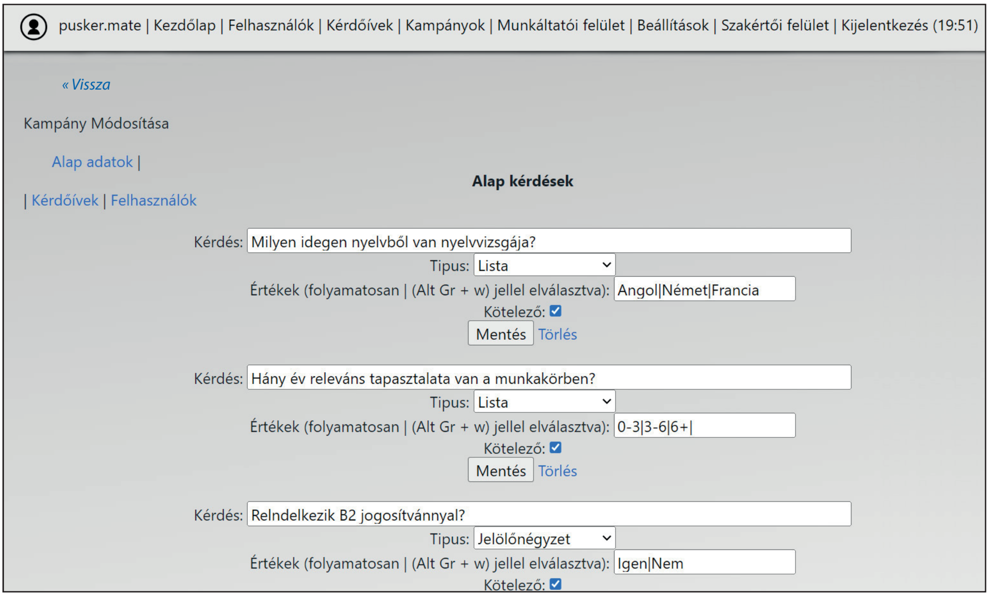
| Felhasználók

Kérdőívek

☐ Kockázatvállalás Kérdőív
☑ RMMT
☑ Rövid Stressz Kérdőív
☑ Munkahelyi elégedettség
☑ Aspirációs Kérdőív
☐ Well Being
☐ Alkalmasságvizsgálat IV.
☐ CIVIL Cégcsoport

Mentés

Appendix 6.



pusker.mate | Kezdőlap | Felhasználók | Kérdőívek | Kampányok | Munkáltatói felület | Beállítások | Szakértői felület | Kijelentkezés (19:51)

« Vissza

Kampány Módosítása

Alap adatok |

| Kérdőívek | Felhasználók

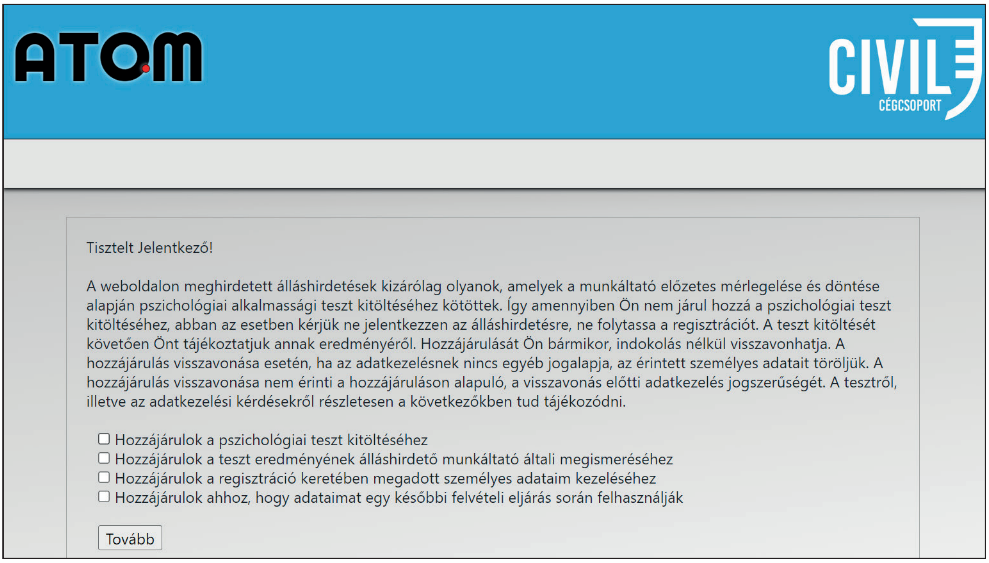
Alap kérdések

Kérdés: Milyen idegen nyelvből van nyelvvizsgája?
Tipus: Lista
Értékek (folyamatosan | (Alt Gr + w) jellel elválasztva): Angol|Német|Francia
Kötelező: ☑
Mentés Törlés

Kérdés: Hány év releváns tapasztalata van a munkakörben?
Tipus: Lista
Értékek (folyamatosan | (Alt Gr + w) jellel elválasztva): 0-3|3-6|6+|
Kötelező: ☑
Mentés Törlés

Kérdés: Relndelkezik B2 jogosítvánnyal?
Tipus: Jelölőnégyzet
Értékek (folyamatosan | (Alt Gr + w) jellel elválasztva): Igen|Nem
Kötelező: ☑

Appendix 7.

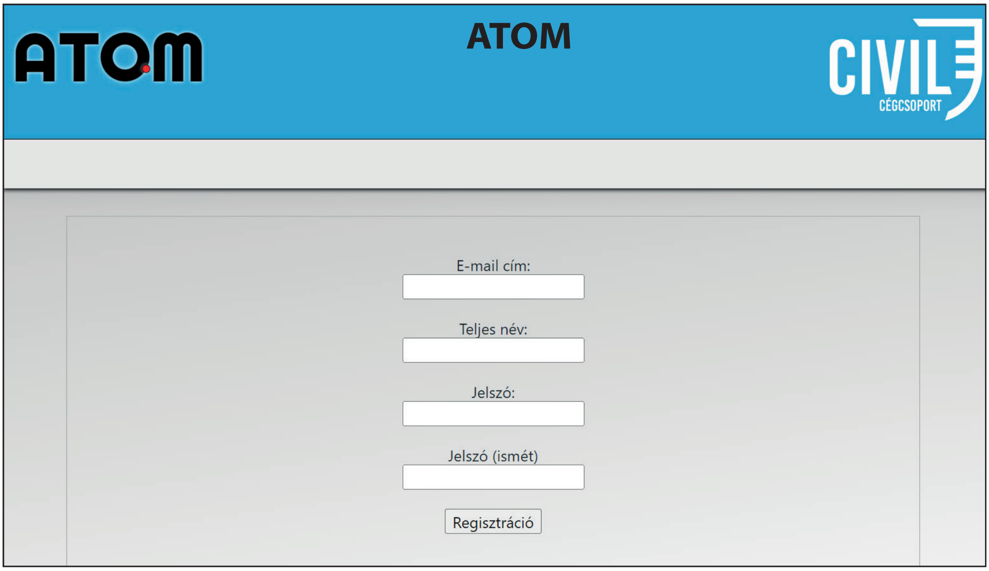


Tisztelt Jelentkező!

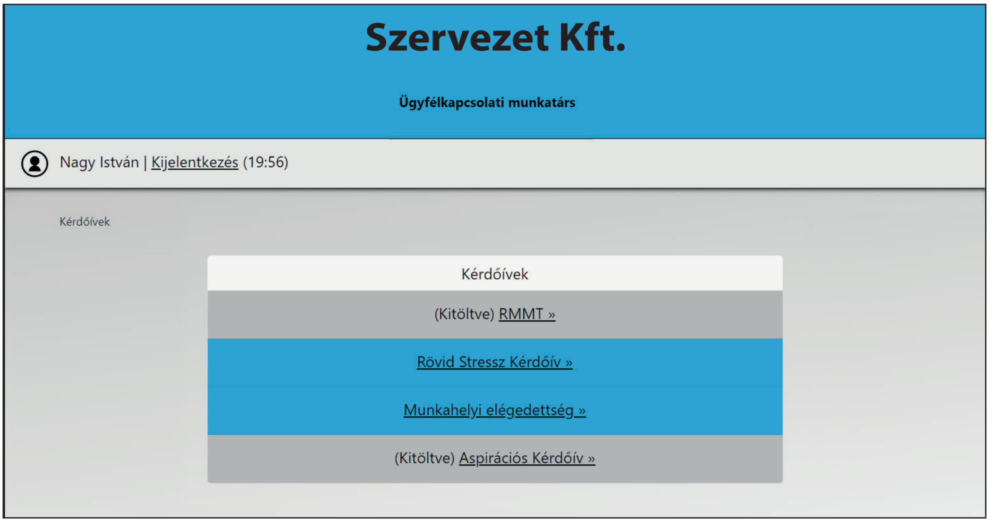
A weboldalon meghirdetett álláshirdetések kizárólag olyanok, amelyek a munkáltató előzetes mérlegelése és döntése alapján pszichológiai alkalmassági teszt kitöltéséhez kötöttek. Így amennyiben Ön nem járul hozzá a pszichológiai teszt kitöltéséhez, abban az esetben kérjük ne jelentkezzen az álláshirdetésre, ne folytassa a regisztrációt. A teszt kitöltését követően Önt tájékoztatjuk annak eredményéről. Hozzájárulását Ön bármikor, indokolás nélkül visszavonhatja. A hozzájárulás visszavonása esetén, ha az adatkezelésnek nincs egyéb jogalapja, az érintett személyes adatait töröljük. A hozzájárulás visszavonása nem érinti a hozzájáruláson alapuló, a visszavonás előtti adatkezelés jogszerűségét. A tesztről, illetve az adatkezelési kérdésekről részletesen a következőkben tud tájékozódni.

☐ Hozzájárulok a pszichológiai teszt kitöltéséhez
☐ Hozzájárulok a teszt eredményének álláshirdető munkáltató általi megismeréséhez
☐ Hozzájárulok a regisztráció keretében megadott személyes adataim kezeléséhez
☐ Hozzájárulok ahhoz, hogy adataimat egy későbbi felvételi eljárás során felhasználják

Appendix 8.



Appendix 9.



Szervezet Kft.

Ügyfélkapcsolati munkatárs

Nagy István | Kijelentkezés (19:56)

Kérdőívek

Kérdőívek

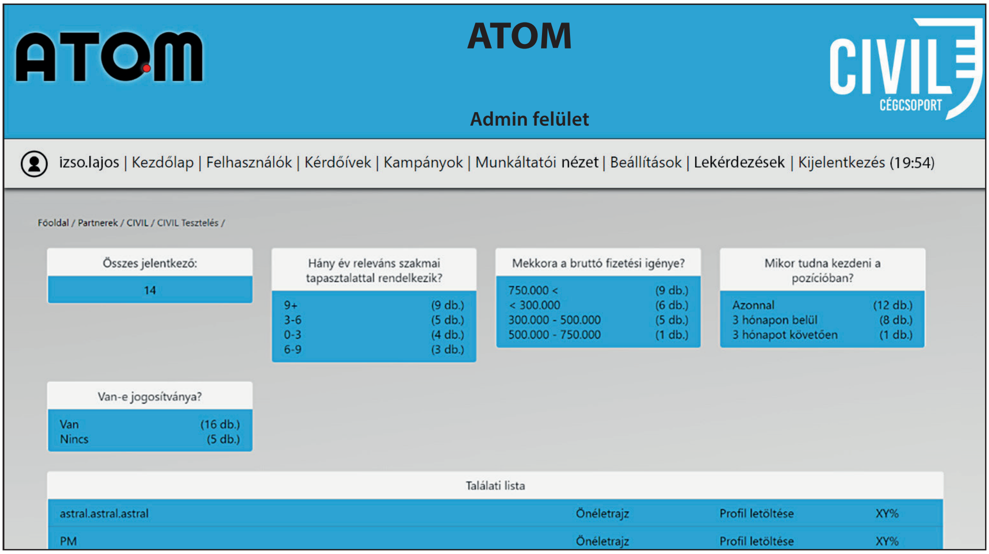
(Kitöltve) RMMT »

Rövid Stressz Kérdőív »

Munkahelyi elégedettség »

(Kitöltve) Aspirációs Kérdőív »

Appendix 10.



ATOM

Admin felület

izso.lajos | Kezdőlap | Felhasználók | Kérdőívek | Kampányok | Munkáltatói nézet | Beállítások | Lekérdezések | Kijelentkezés (19:54)

Főoldal / Partnerek / CIVIL / CIVIL Tesztelés /

Összes jelentkező:
14

Hány év releváns szakmai tapasztalattal rendelkezik?	
9+	(9 db.)
3-6	(5 db.)
0-3	(4 db.)
6-9	(3 db.)

Mekkora a bruttó fizetési igénye?	
750.000 <	(9 db.)
< 300.000	(6 db.)
300.000 - 500.000	(5 db.)
500.000 - 750.000	(1 db.)

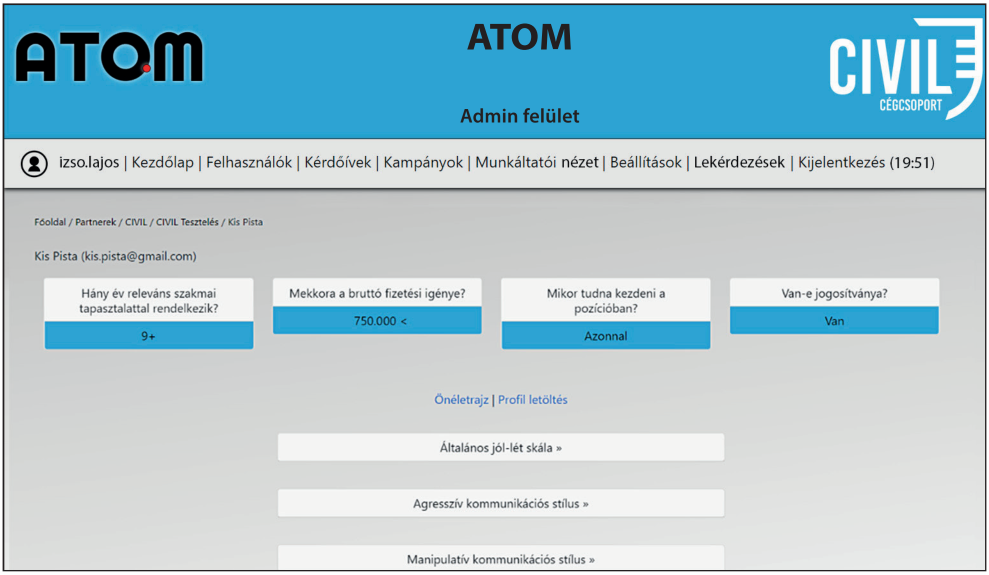
Mikor tudna kezdeni a pozícióban?	
Azonnal	(12 db.)
3 hónapon belül	(8 db.)
3 hónapot követően	(1 db.)

Van-e jogosítványa?	
Van	(16 db.)
Nincs	(5 db.)

Találati lista

astral.astral.astral	Önéletrajz	Profil letöltése	XY%
PM	Önéletrajz	Profil letöltése	XY%

Appendix 11.



ATOM

Admin felület

izso.lajos | Kezdőlap | Felhasználók | Kérdőívek | Kampányok | Munkáltatói nézet | Beállítások | Lekérdezések | Kijelentkezés (19:51)

Fóoldal / Partnerek / CIVIL / CIVIL Tesztelés / Kis Pista

Kis Pista (kis.pista@gmail.com)

Hány év releváns szakmai tapasztalattal rendelkezik?
9+

Mekkora a bruttó fizetési igénye?
750.000 <

Mikor tudna kezdeni a pozícióban?
Azonnal

Van-e jogosítványa?
Van

Önéletrajz | Profil letöltés

Általános jól-lét skála »

Agresszív kommunikációs stílus »

Manipulatív kommunikációs stílus »

JOINTLY APPLYING A WORK SIMULATOR AND ATOM TO PREVENT OCCUPATIONAL ACCIDENTS AND MSD THROUGH WORKFORCE SELECTION

Lajos Izsó
Budapest University of Technology and Economics
izso.lajos@gtk.bme.hu

Blanka Berényi
Károli Gáspár University of the Reformed Church in Hungary
berenyi.blanka.pszi@gmail.com

Máté Pusker
CIVIL Plc
mate.pusker@gmail.com

Summary

Background and Aims: The goal of our work is the presentation of a particular – scientifically well-established – concept aiming to predict the propensity of individual job candidates for causing or suffering workplace accidents, and also for MSD-type *(Musculoskeletal Disorder)* occupational diseases, by further processing the performance parameters obtained by a work simulator (like ErgoScope) with the help of ATOM.

Methods: After introducing the problems of workplace accidents and MSDs, and critically reviewing the basic literature related to the so-called "work sample tests" and work simulators, the application possibilities of a specific, general-purpose work simulator, the ErgoScope, are presented for our purposes. After that, the possibilities of adequately integrating the ErgoScope and ATOM are described with particular respect to workplace accidents and MSDs, illustrated through a fictitious but realistically specified example.

Conclusions: The purposeful combination of the ErgoScope work simulator with ATOM can have a "synergistic" effect that reinforces each other's effects, contributing to a significant reduction in the likelihood of workplace accidents and MSDs. Simply put, we propose to apply the appropriate outputs of the ErgoScope work simulator as inputs to ATOM.

Keywords: workplace accidents, occupational illnesses, MSD, work sample test, work simulator, workforce selection, ErgoScope, ATOM

DOI: 10.17627/ALKPSZICH.2023.3.75

The objectives of this article

The central message of this article is that – if the necessary methodological care is provided – the further processing of performance parameters obtained by a work simulator, with the help of ATOM, might result in a much better prediction of the propensity of individual job candidates for causing or suffering workplace accidents, and also for developing MSD-type *(Musculoskeletal Disorder)* occupational diseases.

The objective is to present this concept shortly, but still as informatively as possible. The tools to achieve this are the following.

- Providing an introduction to the problems of workplace accidents and MSD in Europe, since this is the application field of the proposed approach.
- Giving a concise review of the psychometric properties of work sample tests, since these form the basis for work simulators as workforce selection tools.
- Showing the main functionalities of work simulators (including the ErgoScope) tangentially.
- Having made these preparatory steps above, working out a fictitious, but realistic OSH (Occupational Safety and Health) example for the combined use of the ErgoScope and ATOM.
- Finally, using this fictitious example as a model, discuss the further possibilities and limits of this concept.

The following sections correspond to these points, while in the concluding discussion, an attempt is made to build a scientifically well-established construct for the concept of applying appropriate outputs of the ErgoScope work simulator as inputs to ATOM. Here we will argue why it is worth applying this concept in practice, and some of our related short-term plans are also outlined.

Introduction to the problems of workplace accidents and MSD

Regarding the recent statistics on workplace accidents in Europe (EU-28), EUROSTAT Statistics Explained (2022) provides the following rather gloomy data. In 2020

a) there were 2.7 million non-fatal accidents that resulted in at least four calendar days of absence from work;

b) there were 3,355 fatal accidents (about 20% of them within the construction sector);

c) 44.1% of all non-fatal accidents, and 63.1% of all fatal accidents happened in construction, transportation and storage, manufacturing, agriculture, forestry and fishing sectors;

d) about 66.5% of the total non-fatal accidents involved men;

e) the two types of particularly common injuries were wounds and superficial injuries (26.8% of the total).

Regarding the state of affairs in the field of work-related MSDs in Europe (EU-28), the European Agency for Safety and Health at Work (2023) provides the following shocking pieces of information. MSDs are the most prevalent occupational disease at the European level. Data from self-reporting through surveys *(European Working Condition Survey, European Health Interview Survey, European Labour Force Survey, European Survey of Enterprises on Emerging Risks)* inform us about the following: (1) three out of every five workers complained of MSDs; (2) more often than not, MSDs are accompanied by other health problems; (3) more than a third of workers reported that their work affects their health negatively; (4) 60% of workers with work-related health problem mentioned MSDs as most serious; (5) MSD prevalence is higher among older workers, (6) MSD prevalence decreases with education level.

With the rapid spread of modern information and communication technologies, mental work has become the main field of work for psychological and ergonomic research, while research on physical work has temporarily been neglected. Although the proportion of occupations requiring classic heavy physical work (e.g., miner, loader, earth-worker, material handler, etc.) has decreased radically, this still leaves a smaller number of such jobs. On the other hand – somewhat unexpectedly, and specifically in connection with IT-related jobs – it turned out that even office occupations, traditionally thought of as easy work, can often be physically demanding.

Not only high physical exertions may cause health risks, but also certain physical arm or hand operations – that in themselves, performed only once or several times can be considered as "easy" –, however, if repeated for a large number of times (up to tens of thousands during a work shift!). Examples include prolonged use of a computer keyboard, repeated execution of assembly sub-operations, frequent hand bending, gripping, twisting, squeezing, etc.

Because of the above, in many workplaces, instead of or in addition to the usual performance criteria related to success in the job (work performance), it is reasonable to raise "accident-free" and/or "MSD-free" work to the level of the performance criterion utilizing some reasonable quantification.

Work sample tests

Matching the most important characteristics of a person and a job is essential for job satisfaction and work efficiency. To ensure this, work and organizational psychology have developed several workforce selection approaches. One of them is the simulation of various work situations according to some critical, selected aspects, during which the behaviour of the candidate applying for the given job, is observed and evaluated in a standardized way. This is the work sample test. The advantages and disadvantages of applying such work sample tests are excellently summarized by HR-GuideSurvey.com (2023, opening screen of the link), therefore below we quote its most important parts:

- Main advantages: high reliability and content validity, difficulty for applicants to fake, and use of the same or similar equipment that is used on the job.
- Main disadvantages: costly to administer; and have less ability to predict performance on jobs where tasks take longer time (days or weeks).

Schmidt and Hunter (1998), and later Roth, Bobko and McFarland (2005) carried out large-scale meta-analyses on work sample test validity, and they found that compared with other of the studied procedures for predicting job success, the highest reported validity was for work sample tests (work simulators were not studied directly). The studied procedures were, among others and in increasing order of corresponding biases: work sample tests, integrity tests, conscientiousness tests, employment interview (structured), employment interview (unstructured), job knowledge, peer ratings, reference checks, job experience (years), biographical data measures, ACs (assessment centres), years of education, interests, age, etc.

These findings support the idea, that the use of work sample tests, including work simulators, is a good solution – despite its relatively high cost – in all areas where the consequences of wrong selection decisions could be quite serious. As Izsó (2012) put it, the jobs in which the risk of occupational accidents and MSDs is high, definitely can be considered such an area.

It has to be mentioned, that the ACs also operate partly on the work sample (work simulation) principle, but usually without using simulators, as special hardware and/or software equipment. An AC is a process, where candidates are given carefully designed specific tasks – mainly in the form of work samples – and are evaluated on their ability to perform a particular job. The ACs are strictly job-specific, for example, the book of Hale (2010) is about ACs specifically for selecting police and fire personnel.

Special target devices – called work simulators, to be defined in the next section – can create such simulated work situations of higher fidelity. In addition to the initial application of work simulators for purely aptitude testing (that is, for diagnostic purposes), the use of these devices recently also appeared for developing and improving job-relevant skills (that is, for training purposes).

Work simulators

While simulation, in general, is the imitation of a situation, environment, procedure or process, the simulator is a target device suitable for implementing the simulation itself.

Ergonomics, which according to its brief definition, is a human-centred technological design, deals with the optimization of different Man-Machine-Environment (MME) systems (Hercegfi & Izsó, 2007).

By definition, *ergonomic pathological factors* result from the structure of the MME system, the specific nature of the flow (exchange) of material, energy and information between man and machine, and also between man and the environment, as a result of physical, mental or emotional stress on the person as pathological effects of stress. MSDs are largely caused by *ergonomic pathological factors*, additional basic knowledge about this topic can be found in the publications of Béleczki et al. (2010) and Izsó (2011).

One particular type of simulation is in which a real, "flesh and blood" human gets into interaction with the Machine and/or Environment subsystems of a particular MME system. In what follows, we only deal with such simulations and the related simulators that by definition, are called "work simulators".

A good work simulator behaves largely similarly to the corresponding real Machine objects in terms of essential characteristics when interacting with humans. The degree of this similarity is characterized by *fidelity (realism).*

The *fidelity* of a work simulator, according to its general definition, is the measure of the accuracy of the simulation implemented with the given simulator. It measures how closely the given device follows the evolution (i.e., the behaviour) of the simulated situation, environment or process over time. One of the first reviews of terms, definitions and concepts related to the fidelity of practical work simulators was carried out by Hays (1980), who found that the use of the term (or wording) was not entirely uniform. He found a relative agreement that there are three main types or aspects of simulation fidelity:

- fidelity in external (physical) appearance (at the highest level, "photorealistic");
- functional fidelity (based on a model and related to operation/behaviour);

- psychological fidelity (refers to the sense of reality).

The work simulators operate on the work sample principle and are used mainly for workforce selection and training purposes. Although selecting candidates for given jobs based on their work sample performance goes back many centuries, or even a millennium, the first well-documented and systematic application of this principle is attributed to Hugo Münsterberg. In 1912, he successfully used a railway simulation method for selecting trolley operators first in Boston, and later on in other cities in the USA. Since then, the selection method of simulated work tasks (work sample) quickly spread.

Work simulators appeared in aviation as early as the 1930s. The Link Trainer was one of the first flight simulators, a very simple mock-up plane, designed to train pilots to operate basic flight controls. This later was followed by more and more sophisticated flight simulators, and nowadays already the big majority of civil and military plane types have their own high-fidelity, full-scale training simulators.

Another pioneer area in applying work simulators was the nuclear power industry. By the 1970s fully functional control room simulators had been developed for the most important reactor plant types of that time. The interested reader can find many details on this topic in the book of Skjerve and Bye (2011). The first author of this article has also been involved in developing simulator training methods for the nuclear power industry: Antalovits and Izsó (1999; 2003), Izsó (2001).

In the last several decades, many other vehicles, heavy machine, construction/mining equipment etc. simulators have been developed (e.g., CKAS [2023], TECH-LABS [2023], Caterpillar Inc. [2023], THOROUGHTEC Simulation [2023], CMLABS [2023]), not to speak of sophisticated simulators for military training purposes.

The best way to focus on our main interest presently, the general-purpose work simulators capable of assessing physical abilities, is to refer to the meta-analyses of Gouttebarge et al. (2004). These authors conducted their systematic literature search targeting the four most widely used work simulators (Blankenship system, Ergos work simulator, Ergo-Kit and Isernhagen Work System) in five databases (CINAHL, Medline, Embase, OSH-ROM and Picarta) using the keywords “functional capacity evaluation”, “reliability” and “validity”. They found that although the interrater reliability and predictive validity of the Isernhagen Work System were evaluated as good, the evaluation procedure used was not rigorous enough to allow any valid conclusion. Concerning the other three tools, neither convincing validity nor reliable data were found. These authors concluded that more rigorous studies are needed to demonstrate the reliability and validity of these work simulators.

Since another important work simulator, the Baltimore Therapeutic Equipment (BTE) had been left out from the review by Gouttebarge et al. (2004), a short evaluation of it will be presented here separately. The first important publication on the reliability and validity of BTE came out already more than three decades ago. The authors of this article – Kennedy and Bhambhani (1991) – determined the test-retest reliability and criterion validity of the BTE in an experiment involving 30 male volunteers. These volunteers acted as warehouse goods loaders and performed real (criterion) and simulated handwork. The three criterion tasks were done at light (CL), medium (CM), and heavy (CH) levels of intensity and three corresponding simulated tasks also were done at these three levels of intensity (SL, SM, SH). All of these tasks were repeated in a subsequent session. The authors experienced significant test-retest reliability concerning the two selected physiological parameters (oxygen consumption and heart rate). Although criterion-simulation correlation coefficients were also significant, consistently high criterion validity was found only at CL-SL (for oxygen consumption *r* = .81 and .83; for heart rate, *r* = .88 and .95).

Later some additional important details were published on the reliability and validity of BTE, e.g., Bhambhani, Esmail and Brintnell (1994), Ting et al. (2001).

The Ergoscope work simulator

The general-purpose ErgoScope work simulator, a new Hungarian development, is fitting to the progressive line of the Blankenship system, the Ergos work simulator, the Ergo-Kit, the Isernhagen Work System, and the Baltimore Therapeutic Equipment, and is free from some limitations of these antecedents. The ErgoScope shares the highest similarity with the Ergos and the Baltimore Therapeutic Equipment. The reason why the ErgoScope has been developed was mainly practical: the possibilities of taking into account special domestic needs, availability of quick and flexible service when needed, and detailed documentation in Hungarian.

As with all work simulators, the ErgoScope also simulates the "Machine subsystem" of the MME system corresponding to various work processes and activities. During ErgoScope simulator sessions, essential conclusions can be drawn about the observed person's physical, perceptual-thinking, and – by observation, to a limited extent – emotional characteristics too. *Figure 1* shows that the ErgoScope equipment consists of three standalone workstations (so-called panels) with different functions, which can be operated independently.

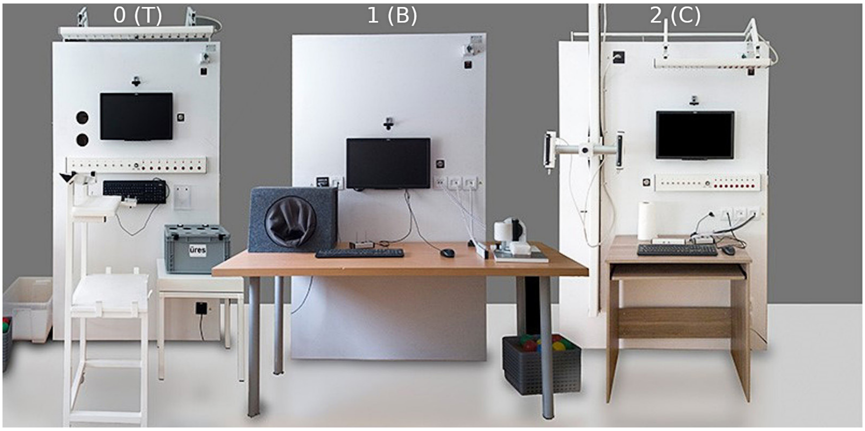


Figure 1. The three panels of the ErgoScope work simulator
Source: https://www.innomed.hu/munkaszimulatorok/

Tasks are performed using various measuring devices connected to data-collecting units, which transmit the measured data to a built-in computer that processes these data.

Panel 0 (T): *Static and dynamic force measurements* (using a bracket, movable on a vertical path):

- Static force measurements (static push/pull horizontally/vertically with two hands)
- Dynamic strength measurements (dynamic lifting to chair/shelf height with two hands, tools for dynamic measurement: scales, chest with weights)

Panel 1 (B): *Examining work performed while sitting*:

- Measurement of grip strength (fist grip with right/left hand, key grip with fingers of right/left hand, 3-point grip with finger of right/left hand, wrist flexion/extension/pronation/supination with right/left hand)
- Touch (with right/left hand)
- Keyboard operation (with right/left hand, with two hands with right/left sign)
- Pencil use ("pencil" use with right/left hand)

Panel 2 (C): *Examining capacity and monotony tolerance* (supplemented by examination of turning, switching and button pressing at chest height and overhead):

- Work capacity (moving crates, sorting balls, rolling balls)
- Monotony tolerance (tray moving, ball sorting, tray scrambling)
- Rotation (rotating knobs with the dominant hand from the eyes/overhead)
- Use of switches (use of switches face-to-face/overhead)
- Use of push-buttons (use of push-buttons directly/overhead)

Altogether 215 concrete-specific objective performance parameters can be measured on these three panels in 36 measurement modes (elementary simulated work situations). The following two main types of work diagnostic surveys are distinguished:

- Full job diagnostic survey: for career guidance, this survey is usually carried out when the client has no concrete ideas about his future job or has several competing ideas but cannot choose. We can often offer the client jobs in which they could perform exceptionally or at least well in terms of the skills required for that job. In the cases of weaker performances, we can recommend targeted skill development. If development is not possible, the search for other, better-fitting jobs follows.
- Targeted job diagnostic survey: this survey is usually carried out when the client comes with a specific job idea or the future employer requests objective data about the client's applicability.

Izsó, Székely and Dános (2015) studied the specific possibilities of this work simulator, especially for use as aptitude testing of people with altered workability, and also touched on its skills development possibilities. (Izsó [2015] compiled a methodological manual, in which the recommended reference values for the ErgoScope parameters can be found.)

Various forms of occupational accidents, MSDs resulting from incorrect/inaccurate limb and full-body movements, and inappropriate exertion can be prevented by properly using the ErgoScope as a professional aptitude testing tool. The parameters measured by the ErgoScope can only be used as predictors of successful future work – i.e., being free from workplace accidents and MSDs – if:

1. these can be considered relevant for the given job based on the knowledge and experience of OSH specialists;
2. a suitable database is available for these parameters for reference purposes.

Regarding the first question, the answer is based on the expertise of OSH specialists. Regarding the second one, we currently have an ErgoScope database, built from the measured parameters of 297 healthy and 100 disabled people. Since installing the first pieces of ErgoScope in 2016, we have participated in many related projects and gained considerable experience in application methodology. The use of ErgoScope parameters as input data (predictors) for the ATOM software package is very promising. As the central message of this article, this problem area is outlined separately in the next section.

Applying data obtained by the Ergoscope as inputs to ATOM for reducing the risk of workplace accidents and MSDs

Requirements for properly combining the ErgoScope and ATOM

As described in more detail in the preceding articles of this special issue – Izsó; Gergely and Takács; Pusker, Gergely and Takács – ATOM, developed by us, is an AI-based expert system for predicting job success based on suitable *predictors* and relevant *success criteria* of the given the job.

A predictor in this context is a variable suitable to predict the future job success of applicants.

Predictors can typically include, among many others: qualifications, relevant work experience, job-specific skills (e.g., driving license, computer proficiency, ability to speak a particular language), certain test scores, objective parameters measured by electro-mechanic or computerized aptitude-testing devices or work simulators, etc.

> Important comment: Since MSDs and occupational accidents often stem from false/imprecise limb and whole-body movements or inappropriate strength exertions, OSH professionals must take into consideration this viewpoint while selecting ErgoScope performance parameters as predictors for a given job.

The *job success criteria* – again, among other things – can typically be:

- actual quantitative and/or qualitative production data (however, such data – for theoretical or practical reasons – are not available for many jobs);
- management's scores on the employee's performance (the disadvantage of these is that they are generally not statistically reliable enough, primarily due to the so-called "halo effect" and the "leniency" and "severity" biases).
- Important comment: in the following fictitious OSH example the long-term accident or/and MSD-freeness must be properly operationalized (quantified) to be used as job success criteria.

If such criteria are available and appropriate – strongly correlated – predictors can also be found for them, based on these predictors, the person's success in the given job can be predicted with a high probability.

ATOM can be applied if valid predictors are available for at least about 100 employees who have already proven to be successful in a given job to different extents (including also failure). This rough practical rule of thumb of using minimally about 100 data points, is based on our experiences gained during targeted ATOM studies.

Providing the required data ATOM's competing and flexible learning algorithms "learn" the relationships between the predictors (as input variables) and the job success criteria (as output variables). Based on the resulting model, ATOM can predict the (expected) success of new applicants for the given job under investigation based only on the values of the predictors (Gergely & Takács, this special issue).

Regarding jobs where the risk of workplace accidents or/and MSD is high, the professionally correctly designed combination of the two domestically developed leading technologies (the ErgoScope with its broad range of functions and the resulting potentially high content validity, and ATOM with its extreme flexibility) is expected to have a strong synergistic effect. This process can be formulated generally with the following simple "IF X, THEN Y" type rule.

1. IF, for a specific job, the OSH specialists determine the parameters (predictors) that can be measured by the ErgoScope and are considered relevant concerning the risk of workplace accidents, or MSD;
2. and these predictors are measured with the help of the ErgoScope in the cases of at least about 100 employees already working in the given position, whose job success in terms of being free from workplace accidents and MSDs are known and numerically different;
3. and after that, the experts use the ATOM's machine learning (ML) algorithms to build the model that best describes the relationship between predictors and job success based on this data;
4. THEN, examining the candidates newly applying for the given job by the ErgoScope, the future job success of these applicants, defined as being accident- and/or MSD-free, can be predicted with relatively high accuracy.

A fictitious OSH example for combining the ErgoScope and ATOM

A goods loader (counter loader) fills shelves and loading areas and keeps the goods clean and tidy in grocery stores, shops, and other wholesale units. For this kind of job, the following ErgoScope performance parameters can be considered relevant:

- Panel 0 (T): static pull / push horizontally / vertically / dynamic lift to chair height/ shelf height;
- Panel 2 (C): work endurance (complex task sequence: moving crates, sorting balls, rolling balls), monotony tolerance (complex task sequence: moving trays, sorting balls, rolling balls).

Let us assume that the ErgoScope performance parameters given above (as predictors) are available for 150 successful and 50 unsuccessful workers for this job, as well as the degree of their actual job success on a five-point scale (its value is 2, 3, 4 or 5 for the successful and 1 for the unsuccessful). By loading this data into ATOM, the learning algorithms "learn" the relationships between the predictors and the job success value given on this five-point scale (in this case, characterizing the persons' middle- or long-term accident or/and MSD-freeness). Finally, if the company wants to hire new employees for this job, then the same ErgoScope performance parameters must be measured for these applicants. Having done so, using the model ATOM provided – based on the data of above mentioned 150 successful and 50 unsuccessful workers – ATOM can predict the probabilities of new applicants falling into each job success category (these are the so-called "labelling probabilities"). Job success (in our particular case, the persons' accident or/and MSD-freeness), however, is also determined by certain psychological characteristics in addition to the physical (motor and force exertion-related) skills/ abilities identified by the OSH professional and measured by the ErgoScope. Therefore, the results of appropriate psychological tests must also be used as predictors, but we will not deal with this problem here (only briefly in the *Discussion*).

In short, in our example, by entering the predictors (the corresponding ErgoScope performance parameters) for the new applicants into ATOM, we get the probabilities with which applicant may fall into each of the five categories of the applied job success scale.

The way of applying ATOM's results depends on the current labour supply and demand situation. When there is labour oversupply, applicants must be sorted according to the decreasing "expected probability of job success" first within the 5 (best) success grades. However, when there is a labour shortage (i.e., when, in principle, all applicants should be hired), hiring those with a 1 (worst) expected job success among new applicants is not recommended. The experience is that these people mainly result an extra expenditure for the company as they eventually either quit on their own or have to be fired.

However, if for some reason, the company is still forced to hire from among the applicants rejected by ATOM, then the applicants must be sorted according to the increasing "probability of job success" first within those with 1 (worst) job success grade. After that, applicants have to be selected from among those, who are relatively lower on this list, and they have to be assessed by traditional HR methods (job interview, overall impression shown at the interview, performance at previous workplaces, living and housing conditions, family situation, the orderliness of finances, etc.). Based on these, the company may override the results obtained from ATOM, but it has, of course, certain risks.

A rule of thumb is that ML models', including ATOM's, classification performance is acceptable if their hit rate is at least about 20% better than the hit rate by chance alone. The background of this guideline, for reasons of space, is not presented here. The actual hit rates (both overall, relating to all categories simultaneously, and also corresponding only to particular categories) can be calculated from the ATOM's output files using the appropriate functionalities of the *Setup (Beállítások)* primary window (Pusker, Gergely & Takács, this special issue, *The four primary windows*).

In our fictitious example, we used a five-point job success scale, to which 1/5 = 0.2 → 20% random hit probability would correspond. If instead, ATOM provides a global forecast with at least 40% accuracy concerning all categories, that is already a significant surplus. However, if our success scale had only two levels (e.g., 0 = *"likely to fail"* and 1 = *"likely to succeed"*), then 1/2 = 0.5 → 50% would be the random hit probability, and this should be increased to at least 70%.

Presenting the proposed approach

During the workforce selection process, the following two main biases are distinguished usually. The first comes from the applicants' side, who are generally willing to pretend to be better than they actually are. This tendency, as Henle et al. (2019) published, might result in faked personality inventories and intentional fraud causing misinterpretation of resumés by HR personnel. The second concerns the applied methods' side – e.g., König and Langer (2022) – since most selection methods involve human decisions, usually by HR personnel, that are inherently error-prone.

We, however, claim that there also exists a third main bias. The source of this relates to the basic question *"Have we chosen the best data processing procedure from among the many possible ones in terms of given input-output relationships?"*. This bias does not relate to data but stems from the chosen data analysing methods.

The first main bias remains henceforward also in the cases of AI-supported workforce selection methods, like ATOM, while the second one can be reduced by applying appropriate AI-driven methods. Reducing the third main bias, however, is only possible if a proper variety of procedures are used, either sequentially or simultaneously, and the results of the best-performing one are accepted. Although this approach requires increased computational resources, it is already quite feasible using today's quick computers. Notwithstanding, we have not found in the literature AI-based methods operating on this principle. Our ATOM system, however, is based on this novel principle: it runs simultaneously many ML algorithms and the outputs of the "winner" (the best performing one) are considered as results (for more details refer to Gergely and Takács: this special issue). The main advantage of competing algorithms is that they can adapt to the diversity of workplace selection, training data of varying size and quality, expert evaluation, and the specific characteristics of the job and latent data generation processes. Thus, our ATOM system can effectively reduce this third type of distortion too. Furthermore, if ATOM uses properly chosen outputs of the ErgoScope work simulator as predictors, this combination hopefully results in relatively bias-free predictions (refer to *Table 1*).

Table 1. A conceptual comparison of hypothesized resultant biases for different combinations of data gathering and data analysing procedures for job success prediction (provided that a simple additive summation rule is valid).

Data analysing procedures	**Data gathering procedures**					
	1. Work sample tests, including work simulators (e.g., ErgoScope)	2. Questionnaire-based methods (conscientiousness tests, integrity tests, etc.)	3. Interview-based methods (structured / unstructured)	4. Peer ratings	5. ACs	6. Biographical measures (years of education / employment, etc.)
1. ATOM	2	3	4	5	6	7
2. Other AI-based methods	3	4	5	6	7	8
3. Traditional statistical methods	4	5	6	7	8	9
4. Traditional non-statistical HR methods	5	6	7	8	9	10

Source: based on own research data

The serial numbers of the data gathering and data analysing methods are at the same time the ranks of the corresponding biases. Thus, concerning data gathering, "work sample tests, etc." the first column has the smallest, while the last column "biographical measures, etc." has the greatest biases. Similarly, concerning data analysis, "ATOM" has the smallest, while "traditional non-statistical HR methods" has the biggest biases. The numeric fields contain the sum of ranks concerning biases corresponding to the procedures in the respective columns and rows. The smaller these sum ranks are, the better (the more bias-free) the corresponding are combinations of the "data gathering" – "data analysis" methods.

The first of the above-mentioned three main biases, attributable to applicants, usually occurs at data gathering procedures 2. and 3. (The biases at procedures 1., 4., 5., and 6. are caused by other factors.)

The second bias, attributable to HR personnel, usually occurs during data gathering procedure 3. (The biases at procedures 1., 2., 4., 5., and 6. are caused by other factors.)

The third bias, attributable to the choice of data processing methods, might occur in all four data analysing procedures, but its magnitude is probably the minimum in the case of ATOM. This is a strong, scientifically well-established argument for using the outputs of the ErgoScope as predictors fed to ATOM.

The accuracy of predictions depends largely on the quality of the input data, as the popular adage says, "garbage in, garbage out". If the algorithms are trained with low-quality data, then the classification result will also be of poor quality. Analyses with low-quality data can raise serious validity problems, but to a certain extent, these can be compensated by using different, more robust statistical procedures (Gergely & Vargha, 2021), as it is done in ATOM.

Since data quality is a multi-dimensional concept, in data science different authors have identified roughly 6–16 distinct dimensions for different purposes. Of these, the first 6 basic dimensions that most publications – e.g., Wang et al. (2002), Batini and Scannapieca (2006), Lee et al. (2006) – are in alignment with. The following two of those are especially relevant to us here: *accuracy* and *relevancy.*

Accuracy: is a measure of how well the data reflects the object being described along the given characteristics (How well does the data reflect reality, irrespective of the relevancy to the actual *matter studied*?). This dimension corresponds to the earlier mentioned first and second main biases.

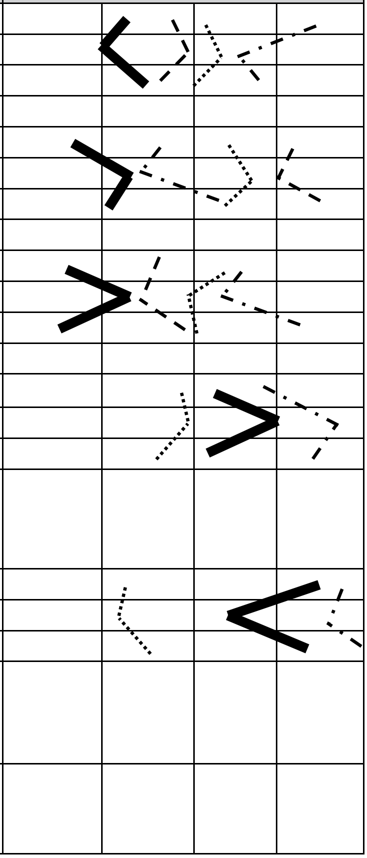
Relevancy: is a measure of the level of consistency between the content of data and the studied areas of interest (in our case, the job success). In other words, it is the extent to which data answers the question of the actual study (To what extent are the data applicable and useful for predicting job success?). Data relevancy means different things for different task contexts: what is relevant for predicting success in a particular job, may not be relevant for other purposes.

Here we go back to the fictional case of selecting candidates for the goods loader job and consider a bit more closely some steps and circumstances of the combined application of the ErgoScope work simulator and ATOM. In the very first step work psychologists and OSH experts – based on their earlier experiences and overall expertise – compile a set of possible predictors consisting of certain personality traits; cognitive, perceptual, motor and force exertion functions. This can be taken as the first iteration step made by human expertise, to be followed by many other computational steps to be done by the concurrent algorithms of ATOM. These starting decisions on the predictors to be applied are decisive since even the best algorithms are later confined by them.

Suppose that the intensity levels of these chosen predictors, minimally necessary for acceptable job performance, are known empirically from the company's earlier workforce selection campaigns. *Table 2* shows this in simplified form: in the "Characteristics" column the chosen predictors are listed, while in the four "Level of Characteristics" columns the minimal requirements are indicated by bold solid polygonal chain lines in percentile units. In the same four columns the actual values of three hypothetical candidates can also be found similarly by dotted, dashed and dotdash chain lines. Suppose that all these data are *accurate* enough.

Table 2. Comparing the fictitious requirements for the goods loader job with the actual values of hypothetical candidate 1, 2 and 3 in terms of competence characteristics.

Group of characteristics	Measuring instruments	Characteristics*	LEVEL OF CHARACTERISTICS (in percentiles**) 0% 25% 50% 75% 100%
Personality traits	Suitable personality tests	Scale 1	
		Scale 2	
		Scale 3	
		Etc.	
Cognitive functions	Suitable cognitive tests	Cogn. function 1	
		Cogn. function 2	
		Cogn. function 3	
		Etc.	
Perceptual functions	Electronic measuring devices	Perc. function 1	
		Perc. function 2	
		Perc. function 3	
		Etc.	
Motor functions**	**ErgoScope work simulator**, special measuring devices	**Moving hutches**	
		Handgrip	
		Wrist stretching	
		Etc.	
Force exertion functions**	**ErgoScope work simulator,** special measuring devices	**Horizontal push**	
		Vertical pull	
		Dynamic lifting	
		Etc.	
Other groups of characteristics as necessary	To be determined…	To be identified…	



Note:

* These characteristics are considered relevant to different degrees for this job. These scales are used as predictors of future job success, and – for simplicity reasons – all are of positive polarity ("the bigger is the better" type).

** A percentile is the percent of cases that are at or below a score.

*** To these functions concrete characteristics (performance parameters) examples are indicated that can be measured by the ErgoScope work simulator.

━━━━━	Requirements by a given hypothetical job
…………………	Values of hypothetical candidate 1
‒ ‒ ‒ ‒ ‒ ‒	Values of hypothetical candidate 2
‒ · ‒ · ‒ · ‒	Values of hypothetical candidate 3

Source: edited on the basis of own research data

The percentiles are proper units for both the minimal job requirement and the actual values of candidates since these correctly reflect the fact that if a predictor value is very infrequent in the population of possible candidates, that very predictor has very high predictive power. This job, as seen in *Table 2*, requires such a high handgrip value that about 75% of the population cannot produce. We can also see that hypothetical candidate 3 is able to exert handgrip that about 90% of the population cannot do. On the contrary, the requirement concerning scale 2 personality trait is only about a 25% percentile, which about 75% of the population can perform.

Concerning hypothetical candidate 1, we can see that while in the cases of personality traits, cognitive and perceptual functions, the values of characteristics are above the minimal requirements of the job, in the cases of motor and force exertion functions the values are below the minimum requirements. This data set is *relevant* since contains the appropriate motor and force exertion characteristics (it is another question that since these characteristics are below the required level, these decrease the success probability in the goods loader job).

Concerning hypothetical candidate 2, we can see that while in the cases of personality traits, cognitive and perceptual functions, the values of characteristics are above the minimal requirements of the job, in the cases of motor and force exertion functions the values are missing. This data set is *irrelevant* since this only contains such characteristics that have little or almost nothing to do with the success of the goods loader job. This fact represents a lack of information concerning the success probability in the goods loader job.

Concerning hypothetical candidate 3, we can see that in the cases of all characteristics, the values are above the minimal requirements of the job. This data set is *relevant* since contains the appropriate motor and force exertion characteristics (and since all these characteristics are above the required, these increase the success probability in the goods loader job).

Psychology, as a pure theoretical science, primarily wants to explain psychic phenomena with the simplest and most parsimonious models possible, while placing less emphasis on prediction. The consequence is that the results can only be generalized within a closed theoretical framework and often have negligible predictive power (Robinaugh et al., 2021). In contrast, ML algorithms (especially deep neural networks) aim to maximize the prediction accuracy of the models, and mostly they do not provide an understandable explanation for how the phenomenon actually works (Yarkoni & Westfal, 2017).

Therefore, when we started developing ATOM for applied work and organizational psychological purposes, at the same time we also decided in favour of maximizing the prediction accuracy, and based on this, maximizing the efficiency of practical workforce selection decisions. The price we have to pay for it is that we will not necessarily know which variables and to what extent played a role in the outcome.

These limitations have certain consequences concerning *Table 2*. This table, as it is, has mainly didactic goals and therefore its content and the related interpretation above are rather simplified.

Although work psychologists and OSH experts naturally can compile a valid set of possible predictors, in reality they can never determine in advance the intensity levels of these chosen predictors, in concrete numerical terms, that are minimally necessary for acceptable performance in a given job. The reason for it is that ATOM's ML algorithms, optimized for maximum prediction accuracy, hardly provide any information about how the predictors are actually interacting (increasing or decreasing each other's effects), and consequently, how much is the resultant predictive power of the individual predictors. So, it could still happen, that a predictor thought rightfully very relevant by a human expert, turns out to be seemingly unimportant because of the confusing complex interactions between the many predictors. This is especially true if the number of predictors is high (say several hundred).

Discussion

From the EUROSTAT Statistics Explained (2022) publication, we have learnt that although in Europe significant progress has already been made in the field of OSH in recent decades, still more than 3,300 fatal and about 2.7 million non-fatal accidents occur in the 28 EU member states every year. The most prevalent occupational diseases still are MSDs: three out of every five workers complained of MSDs in the last years. These facts justify why preventing workplace accidents and MSD-type occupational diseases is regarded as a primary goal nowadays.

As mentioned earlier in the *Work sample tests* section, the meta-analyses on work sample tests revealed that compared with other procedures for predicting job success, the highest reported validity was for work sample tests. Therefore, an effective way for preventing workplace accidents and MSDs could be to develop workforce selection methods targeting specifically these problems based on appropriate work simulators, which are purposeful implementations of carefully selected work sample tests. We can sum up that a professionally appropriate combination of the use of the ErgoScope work simulator and the capabilities of ATOM may result in a "synergistic" effect, reinforcing each other's effects thus contributing to a further reduction in the occurrence of workplace accidents and MSDs. Therefore, applying appropriate outputs of the ErgoScope work simulator as inputs to ATOM is proposed.

In a recent OSH conference – Izsó (2022) – we announced our plan to realize the fictitious example of the goods loader job, discussed above, in the near future in the form of a large-scale field study. Similarly, in the longer term, we also plan to carry out job success prediction studies by the combined use of the ErgoScope and ATOM involving other physically demanding jobs.

Összefoglalás

Munkaszimulátor és az ATOM együttes alkalmazása munkabalesetek és MSD típusú foglalkozási megbetegedések megelőzésére a munkaerő kiválasztása útján

Háttér és célkitűzések: Munkánk célja annak a – tudományosan jól megalapozott – koncepciónak a bemutatása, amely szerint egy adott fizikai munkakörre jelentkező munkavállaló hajlama munkabaleset előidézésére vagy elszenvedésére, illetve MSD *(Musculoskeletal Disorder)* típusú foglalkozási megbetegedésre jól előrejelezhető, ha prediktorként az ATOM rendszerben munkaszimulátorral (pl. ErgoScope-pal) nyert releváns mérési adatokat használunk.

Módszer: A munkabalesetek és MSD típusú foglalkozási megbetegedések problémakörének általános ismertetése, valamint az ún. „munkaminta tesztekkel" és munkaszimulátorokkal kapcsolatos szakirodalom kritikai áttekintése után egy konkrét általános célú munkaszimulátor, az ErgoScope alkalmazási lehetőségeit vizsgáljuk jelenlegi céljaink kapcsán. Ezt követően annak a lehetőségeit vizsgáljuk meg – egy fiktív, de realisztikus példával illusztrálva –, hogy hogyan lehet a legelőnyösebb módon integrálni az ErgoScope és az ATOM rendszereket a munkabalesetek és az MSD típusú foglalkozási megbetegedések lehető legpontosabb előrejelzése érdekében.

Következtetések: Az ErgoScope és az ATOM együttes alkalmazása egyfajta „szinergikus" hatást eredményezhet, amely felerősíti a két rendszer külön-külön történő alkalmazásának a hatásait, és ez jelentősen hozzájárulhat a munkabalesetek és az MSD típusú foglalkozási megbetegedések valószínűségének csökkenéséhez. Egyszerűen fogalmazva, azt javasoljuk, hogy az ErgoScope-pal nyerhető, szakszerűen kiválasztott mérési adatokat (az ErgoScope alkalmas kimeneteit) alkalmazzuk az ATOM rendszer bemeneteiként.

Kulcsszavak: munkabaleset, foglalkozási megbetegedés, MSD, munkaminta teszt, munkaszimulátor, munkaerő-kiválasztás, ErgoScope, ATOM

References of this Special Issue

Izsó, L. (2023). The concept of an AI-based expert system (ATOM) for predicting job success. *Alkalmazott Pszichológia*, *25*(3), 5–13.

Gergely, B., & Takács, Sz. (2023). ATOM – a flexible multi-method machine learning framework for predicting job success. *Alkalmazott Pszichológia*, *25*(3), 15–30.

Pusker, M., Gergely, B., & Takács, Sz. (2023). ATOM's structure – employee and employer feedback, survey site. *Alkalmazott Pszichológia*, *25*(3), 53–72.

References

Antalovits, M., & Izsó, L. (1999). Self-assessment and learning in nuclear power plant simulation training. In J. Misumi, R. Miller, & B. Wilpert (Eds.), *Nuclear Safety: A Human Factors Perspective* (pp. 243–256). Taylor & Francis.

Antalovits, M., & Izsó, L. (2003). Assessment of Crew Performance and Measurement of Mental Effort in a Cognitively Demanding Task Environment. In Hockey, G. R. J., Gaillard, A. W. K., & Burov, O. (Eds.), *Operator Functional State* (pp. 284–290). NATO Science Series, IOS Press.

Batini, C., & Scannapieca, M. (2006). *Data Quality. Concepts, Methodologies and Techniques.* Springer-Verlag.

Béleczki L., Varga J., Plette R., & Ungváry Gy. (2010). Ergonómiai alapismeretek. Ergonómiai kóroki tényezők okozta megbetegedések. In Ungváry Gy., & Morvai V. (Eds.), *Munkaegészségtan. Foglalkozás-orvostan, foglalkozási megbetegedések, munkahigiéné* (pp. 711–718; 3. átdolgozott és bővített kiadás). Medicina Könyvkiadó Zrt.

Bhambhani, Y., Esmail, S., & Brintnell, S. (1994). The Baltimore Therapeutic Equipment work simulator: Biomechanical and physiological norms for three attachments in healthy men. *American Journal of Occupational Therapy, 48*(1), 19–25. 10.5014/ajot.48.1.19. PMID: 8116780, https://www.semanticscholar.org/paper/The-Baltimore-Therapeutic-Equipment-work-simulator%3A-Bhambhani-Esmail/5c787f8937148587c244e4f6ee8d21b81950e67d

Caterpillar Inc. (2023, January 21). *CAT simulators*. https://catsimulators.com/videos

CKAS (2023, January 21). *Driving Simulators*. https://www.ckas.com.au/driving_simulators_29.html?gclid=EAIaIQobChMI4oXbpMf2_AIV0EeRBR1QegLvEAAYASAAEgKH6_D_BwE

CMLABS (2023, January 25). *Construction Equipment Training Simulators*. https://www.cm-labs.com/immersive-simulation-products/construction-equipment-training-simulators/

European Agency for Safety and Health at Work (2023, February 11). *Work-related diseases*. https://osha.europa.eu/en/themes/work-related-diseases

European Agency for Safety and Health at Work (2023, February 10). *Work-related musculoskeletal disorders: prevalence, costs and demographics in the EU.* Publications Office. https://data.europa.eu/doi/10.2802/66947

EUROSTAT Statistics Explained (2022, December 11). *Accident at work statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Accidents_at_work_statistics

Gergely, B., & Vargha, A. (2021): How to Use Model-Based Cluster Analysis Efficiently in Person-Oriented Research. *Journal for Person-Oriented Research*, *7*(1), 22–35.

Gouttebarge, V., Wind, H., Kuijer, P. P. F. M., & Frings-Dresen, M. H. W. (2004). Reliability and validity of Functional Capacity Evaluation methods: A systematic review with reference to Blankenship system, Ergos work simulator, Ergo-Kit and Isernhagen Work System *International Archives of Occupational and Environmental Health, 77*(8), 527–537 https://doi.org/10.1007/s00420-004-0549-7

Hale, C. D. (2010). *The Assessment Center Handbook for Police and Fire Personnel*. Charles C. Thomas Publisher Ltd.

Hays, R. T. (1980). *Simulation Fidelity: A Concept Paper*. U. S. Army, Research Institute for the Behavioral and Social *Sciences*.

Henle, C. A., Dineen, B. R., & Dulffy, M. K. (2019): Assessing intentional resume deception: Development and nomological network of a resume fraud measure. *Journal of Business and Psychology, 34*(1), 87–106.

Hercegfi K., & Izsó L. (Eds.) (2007). *Ergonómia*. BME tankönyv. Typotex Kiadó.

HR-GuideSurvey.com (2023). Personnel Selection: Methods: Work Sample Tests. https://hr-guide.com/Selection/Work_Sample_Tests.htm

Izsó, L. (2001). Fundamentals of the Model Behind the COSMOS Methodology Used for Team Assessment in Simulator Training. *Journal of Occupational Safety and Ergonomics*, *7*(2), 163–178.

Izsó L. (2011): Ergonómiai tényezők vizsgálata a munkahelyen. In Ungváry Gy. (Ed.), *Munkahigiénés gyakorlatok*. OMFI.

Izsó L. (2015): *Az ErgoScope munkaszimulátor által mérhető paraméterek javasolt indulási referencia-értékei*. Methodological Handbook. InnoMed.

Izsó L. (2022): *A munkaképesség felmérése munkaszimulátor alkalmazásával, a beválás előrejelzése mesterséges intelligencia segítségével.* Magyar Üzemegészségügyi Tudományos Társaság XI. Kongresszusa. 2022. október 6. http://www.mutt.hu/anyagok/MUTT_kongresszus_2022_Esztergom_Programfuzet.pdf

Izsó L., (2012): Munkaszimulátorok alkalmazásának lehetőségei a munkavégzés biztonságának javításában. *Munkavédelem* és *Biztonságtechnika*, *2012*(4), 10–16.

Izsó, L., Székely, I., & Dános, L. (2015): Possibilities of the ErgoScope high fidelity work simulator in skill assessment, skill development and vocational aptitude tests of physically disabled persons (*„Best Paper Award”* winner conference paper). 13th International Conference of the Association for the Advancement of Assistive Technology in Europe, Sept. 9–12, 2015. Budapest, Hungary. As book chapter: In Sik-Lányi, C., Hoogerwerf, E. J., Miesenberger, K., & Cudd, P. (Eds.). *Assistive Technology* (pp. 825–831). IOS Press.

Kennedy, L. E., & Bhambhani, Y. N. (1991). The Baltimore Therapeutic Equipment Work Simulator: Reliability and validity at three work intensities. *Archives of Physical Medicine and Rehabilitation*, *72*(7), 511–516.

König, C. J., & Langer, M. (2022): Machine learning in personnel selection. In Strohmeier, S. (Ed.), *Handbook of Research on Artificial Intelligence in Human Resource Management* (pp. 149–167). Edward Elgar.

Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2006), *Journey to Data Quality.* MIT Press.

Robinaugh, D. J., Haslbeck, J. M., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021): Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743.

Roth, P. L., Bobko, P., & McFarland, L. A. (2005): A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*(4), 1009–1037. https://doi.org/10.1111/j.1744-6570.2005.00714.x

Schmidt, F. L., & Hunter, J. E. (1998): The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2). 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Skjerve, A. B., & Bye, A. (2011): *Simulator-based Human Factors Studies Across 25 Years*. Springer.

TECH-LABS (2023, January 30). *Forklift Personal Simulator*. https://tech-labs.com/products/forklift-personal-simulator

THOROUGHTEC Simulation (2023, January 30). *Full mission simulators*. https://www.thoroughtec.com/cybermine-full-mission-mining-simulators/?gclid=EAIaIQobChMIprLdqMr2_AIVCNGyCh28zA9VEAAYASAAEgLWQfD_BwE

Ting, W., Wessel, J., Brintnell, S., Maikala, R., & Bhambhani, Y. (2001). Validity of the Baltimore Therapeutic Equipment Work Simulator in the Measurement of Lifting Endurance in Healthy Men. *American Journal of Occupational Therapy*, *55*(2), 184–190. https://doi.org/10.5014/ajot.55.2.184

Wang, R. Y., Ziad, M., & Lee, Y. W. (2002). *Data Quality.* Kluwer Academic Publishers.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

ILLUSTRATING REAL-LIFE ATOM APPLICATION CASE STUDIES

Lajos Izsó
Budapest University of Technology and Economics
izso.lajos@gtk.bme.hu

Blanka Berényi
Károli Gáspár University of the Reformed Church in Hungary
berenyi.blanka.pszi@gmail.com

Szabolcs Takács
Károli Gáspár University of the Reformed Church in Hungary
takacs.szabolcs.dr@gmail.com

Summary

Background and Aims: Presenting real-life ATOM application field studies to illustrate how ATOM should be applied in the practice of workforce selection.
Methods: After having defined applied metrics for assessing the categorization performance of ATOM, and – for simplicity, reliability and uniformity reasons – confining ourselves to binary job success scales. Five concrete real-life ATOM application field studies are presented basically in tabular form.
Discussion: It can be stated that (1) ATOM is susceptible to data quality, therefore pertinent job success and predictor data are needed; (2) the sample sizes must always be at least about 100; (3) the free choice of cut-off points on the label probability scales, as necessary, is an effective method for finding the best solution.
Keywords: ATOM, recruitment, workforce selection, cut-off levels, categorization performance

Introduction

The proper management of HR (human resources) at working organizations is of decisive importance. The HRM (human resources management) covers the primary fields of recruiting, workforce selection, employment, training, performance monitoring/management, waging, labour relations, and occupational safety and health. This article focuses on selection, which is a decision-making process still made mainly by human personel.

DOI: 10.17627/ALKPSZICH.2023.3.95

Experience shows that humans, like HR persons, usually underperform in workforce selection decisions. Eubanks (2022) states: “Admit it: we’re bad at the selection. The data shows that the common ways we interview and many of the methods companies use to rank candidates (school attended, college grades, or other demographic data) are highly unreliable statistically” (p. 109). An appropriate AI-supported workforce selection method could be free from the serious validity limits of traditional methods described by Barrick et al. (2001) and Henle et al. (2019).

These, and other similar experiences, were strong arguments to us for developing a sophisticated AI application to support workforce selection, called ATOM (Artificial intelligence for Testing Occupational success of Manpower).

The basic function of ATOM is to “learn” the relationship between suitable *predictors* and relevant *success criteria* of the given job. A predictor in this context is a variable suitable to predict the future job success of applicants, while the *job success criteria* can typically be actual quantitative and/or qualitative production data, management’s scores on the employee’s performance, etc.

A novel feature of ATOM is – as described in (Gergely & Takács, this special issue) – that in its core many machine learning (ML) algorithms run concurrently, and the results of the best-performing algorithm are accepted. ATOM works via the type *“supervised learning”* of the ML, where the “training example” is a set of input-output data pairs. The goal of the process is classification, that is, to estimate probabilities for each new candidate falling into different success categories and then, based on these, to determine success categories themselves solely from the predictors.

The reader can find further details about the wider HRM context of ATOM and some basic information on ATOM’s algorithms in (Izsó, this special issue).

The ATOM software package can handle job success data on any type of discrete scale. If job success data are available on other scales in the practice, these must be transformed to a discrete scale before feeding them into ATOM.

This article presents specific ATOM application case studies using ATOM’s experts’ functionalities, but the employees’ and employers’ functionalities were not considered here. However, it should be noted, that the purposeful operation of ATOM in the future should also involve these functionalities. While, in our cases, all the predictors and job success data were entered into ATOM as external files, in the future, the data obtained directly from the employees online (e.g., the completed questionnaires) shall be collected in internal files through the employees’ functionalities. That way, the procedure will be automatic and very quick.

ATOM works with two types of input data files in a predefined specific format, which contains a personal identification code and predictor variables of any number and any scale in addition to discrete-scale job success data. ATOM can handle only one job success variable at a time within one run. So, if we have more than one job success variable, they must be analysed separately.

After running, ATOM provides the results organised into four types of output data files in specific predefined formats. The most important of these results are:

- *Predicted job success categories*, together with the related expected *category probabilities* (called also *labelling probabilities,* the probabilities of falling into each category for each person) are calculated by the "winner" algorithm. ATOM adds a person into specific success category which has the highest category probability calculated by the algorithms of ATOM. Experience provided us with good reasons to analyse these probabilities directly (especially in the case of binary scales) instead of the resulting categories. We follow this path in this special issue while presenting the case studies.

- *Classification table* – also known as *confusion matrix* – is also calculated by the "winner" algorithm, characterises the constructed "winner" model's goodness under the given circumstances.
- Indicators characterising the *predictive power for each predictor,* are calculated by the best-performing logistic regression algorithm. These are the magnitudes and related statistical significance levels of the logistic regression coefficients that best fit the given model.

The results obtained from ATOM have different consequences for practical use if there is an oversupply or an undersupply of the labour force. Therefore, as explained in more detail in (Izsó, Berényi & Pusker, this special issue), the particular way applying ATOM's results fundamentally depends on the current labour force situation.

Applied metrics for assessing the categorization performance of ATOM

The analysis using job success probabilities can often be radically simplified – quite independently of the number of categories of the original job success scale used during actual data collection in the field – by confining ourselves only to two-point (i.e., binary) job success scales (e.g., 0: *"not likely to succeed"*; 1: *"likely to succeed"*). In this case, the analysis can be performed using one single success probability scale; therefore, there is no need for probability analysis to be performed separately for each category.

Besides simplicity, binary success scales are also justified by uniformity and reliability. While uniformity represents only a convenience point of view, the reliability issue has theoretical significance.

As Alwin, Baumgartner and Beattie (2018) put it, measurement results are the most reliable when fewer response categories are used. Thus, binary scales have the highest reliability. On the other hand, response categories of higher numbers may have the advantage that more scale points will capture more variation (which could be critical in doing correlations or regressions). A large part of that variation is, however, as we know from experience, "noise" from measures that become increasingly unreliable.

Reducing more general problems to binaries has one more advantage of making certain concepts, metrics and procedures – developed specifically for binary problems in machine learning (ML) – applicable to ATOM analyses. The four most important, simple and widely used metrics (*overall hit probability, sensitivity, specificity*, and *precision*) and the related procedures (analyses based on *ROC* curves and *Precision-Recall* curves) applicable to assess the categorisation performance of ATOM, are briefly summarised below first by defining them by plain text, later a bit more formally, defining them by formulas too.

1. *Overall hit probability* (called also *overall hit rate* or *percentage of correctly classified cases*) is the overall probability that ATOM will correctly categorize a case.

   As it was pointed out (Gergely & Takács, this special issue), this metric is used as an efficiency indicator of ML algorithms running simultaneously within ATOM. Of the competing algorithms, the "winner" has the highest *overall hit probability.* A *"high enough"* value of *overall hit probability* is only the necessary condition for practical usability. For being *"high enough"* the generally accepted rule of thumb for binaries: anything greater than 0.70 (70%) is *"high enough"*. The sufficient condition, in addition to the necessary condition, is that – depending on the actual goal of analysis – either *sensitivity* or *specificity,* or both, should also be *"high enough"* (*sensitivity* and *specificity* are defined in the following two paragraphs). If *sensitivity* or *specificity* is not *"high enough"*, purposefully selecting another cut-off point – instead of the default 50% – on labelling reliability could improve these metrics, but there is no guarantee for that.
2. *Sensitivity (recall)* is the probability that ATOM will categorize a case as positive that is truly positive.
3. *Specificity* is the probability that ATOM will categorize a case as negative that is truly negative.
4. *Precision (Positive Predictive Value)* is the probability that a case categorized by ATOM as positive is truly positive.

As already mentioned in Izsó (this special issue), similar to the approach by Tasdemir (2015), we use ROC analysis for evaluating ATOM's classification performance, and also as a kind of validity detection.

To make the above a bit more precise and adapted to ATOM, let the following classification table (confusion matrix) be given, where job success is defined on a binary scale, the categories of which are: 1 = *"less likely to be successful in the job"*, 2 = *"more likely to be successful in the job"*. This job success scale will be used uniformly in the following four case studies (in the fifth case study these categories will be related not to job success, but to work motivation).

It has to be emphasised again, that ATOM can process job success data on any type of discrete scales, but in this article, we confine ourselves to binaries. In reality, in these case studies job success data originally were not given on binary scales, but for simplicity and uniformity reasons these all were transformed into binaries. In the 1st, 2nd, 3rd and 4th case studies of this special issue, job success was originally available on a 5-point scale, while in the 5th case study on a 3-point scale.

Table 1. A classification table with general notations for deriving *overall hit probability, sensitivity, specificity* and *precision* metrics

		Categorization by ATOM		
		1 (+)	2 (–)	∑
Actual category	1 (+)	TP	FN	TP + FN
	2 (–)	FP	TN	FP + TN
	∑	TP + FP	FN + TN	TP + FN + FP + TN

Source: edited by using own research data

From now on, by definition, category 1 should be taken as positive (+) in the sense that persons belonging to this category do have a set of characteristics that work against their suitability for the given job.
TP = True Positive = number of cases truly (correctly) categorized by ATOM as positive
TN = True Negative = number of cases truly (correctly) categorized by ATOM as negative
FP = False Positive = number of cases falsely (incorrectly) categorized by ATOM as positive
FN = False Negative = number of cases falsely (incorrectly) categorized by ATOM as negative

Based on the above, the textually introduced four metrics are formally defined in the following way.

1. *Overall hit probability (overall hit rate, percentage of correctly classified cases)* = (TP + TN)/(TP + FN + FP + TN),
   the overall probability that ATOM will correctly categorize a case. This metric is calculated for all competing algorithms by ATOM, and the particular algorithm providing its highest value is considered to be the "winner".
2. *Sensitivity (recall, failure prediction probability)* = TP/(TP + FN),
   the probability that ATOM will categorize a case as positive that is truly positive. Its value, by definition, is 0 if TP = 0 and is 1 if FN = 0.
3. *Specificity (success prediction probability)* = TN/(TN + FP),
   the probability that ATOM will categorize a case as negative that is truly negative. Its value, by definition, is 0 if TN = 0 and is 1 if FP = 0.
4. *Precision (Positive Predictive Value)* = TP/(TP + FP), the probability that a case categorized by ATOM as positive is truly positive. Its value, by definition, is 0 if TP = 0 and is 1 if FP = 0.

These commonly used concepts originally came from chemical analytics and medical diagnostics (e.g., testing the presence of arsenic in drinking water, pregnancy tests, or COVID tests) into the field of ML.

The latest three metrics are not calculated by ATOM itself, but if these are necessary for deeper analysis, these can quickly be calculated with the help of suitable external pieces of software (e.g., Excel, IBM SPSS Statistics, SAS, etc.).

In general, if a job success scale has L categories, the probability that a person falls into a particular success category merely by chance is $p = 1/L$. In the case of binary scales $L = 2$, therefore, the corresponding chance probability is $p_1 = p_2 = ½ = 0,5$ (also called 50%). For a binary job success scale, the category probability p_1 means the probability that a person belongs to success category 1. The related 50% chance probability ($p_1 = 0,5$) is taken by ATOM as the default „cut-off" level, above which the person belongs to success category 1, below which belongs to success category 2. The p_1 and p_2 category probabilities add up to 1: $p_1 + p_2 = 1$.

As we defined category 1 as *"less likely to be successful in the job"*, and category 2 as *"more likely to be successful in the job"*, in this respect p_2 is not just a category probability but also the success probability (while p_1 is the failure probability). Experience has shown that there are situations where using "cut-off" levels other than 50% could provide better results for specific problems.

The actual *overall hit probabilities* based on the default 50% cut-off level, and also those that belong to purposefully selected other particular cut-off probabilities, were calculated from ATOM's output files titled *pred_output.csv* via the appropriate functionalities accessed in the *Setup* primary window (Pusker, Gergely & Takács, this special issue, *The four primary windows*). Additional analyses in these case studies were performed using IBM SPSS Statistics version 28.

The above shows that the default (relating to $p_1 = 0,5$) classification tables can only be interpreted directly to a somewhat limited extent. However, from the corresponding category probabilities, new classification tables can be constructed as necessary, for any other optional cut-off levels, again with the help of suitable external programs (Excel, IBM SPSS Statistics/Modeler, SAS, etc.).

ROC curves are diagrams characterising the performance of a binary categorisation/classification system (in our case, ATOM), which represent *sensitivity* as a function of (1 – *specificity*). In other words, it plots the probability of a *true alarm* (TP) as a function of the probability of a *false alarm* (FP). The curve shows the possible trade-offs between true and false alarms for different *sensitivity (recall)* and *specificity* cut-off levels.

It is important to note that there are two different kinds of cut-off levels used in this article, not to be confused. While the p_1 and p_2 category probabilities provided by ATOM reflect only the uncertainty of categorisation, the *sensitivity (recall)* and *specificity* appearing on the axes of the ROC curves, as defined earlier, are conditional probabilities. Consequently, by changing the cut-off levels of p_1 (or p_2) we can produce new classification tables. By changing cut-off levels of *sensitivity* (or *specificity),* however, we can find different trade-offs on a ROC curve between *sensitivity* and *specificity.*

The great advantage of ROC curves thus is that they simultaneously provide aggregated information about the discrimination performance of the given system for all possible *sensitivity* / *specificity* cut-off levels, compared to e.g. with the different classification tables, all of which only refer to one specific cut-off level of p_1 (or p_2).

At the same time, in the relatively often occurring "imbalanced" samples, in which the number of positive cases is significantly (sometimes even by orders of magnitude) smaller than the number of negative cases, the results obtained from the ROC curves are somewhat distorted. Therefore, the so-called Precision-Recall curves were developed just to analyse such "imbalanced" samples.

Precision-Recall curves are also diagrams characterising the performance of a categorisation/classification system (in our case, ATOM), representing *precision* as a function of *sensitivity (recall).* These curves, however, focus on the cases categorised as positive (in our case category 1), so the potentially large number of actually negative cases does not distort the analysis. Similar to ROC curves, this curve shows the possible trade-offs between *precision* and *sensitivity* for different axis cut-off levels. The interested reader can have further information about ROC and Precision-Recall curves from Davis and Goadrich (2006) and at related links.

An example of interpreting *ROC* curves and *Precision-Recall* curves at varying levels of cut-off points on their axes, can be found later concerning *Picture 1.*

We worked with several "conflicting" questionnaires during the presented case studies. These were competing with each other because some questionnaires were our own developments during the project, so we also included questionnaires that served the convergent and divergent validity of the questionnaire to be developed.

Based on these measuring instruments, we carried out the necessary runs and analyses using the previously described (Pusker, Gergely & Takács, this special issue) questionnaire entry page and ATOM-CORE analyses. After the preparatory phase, we were able to record the different measuring devices on different platforms (for example, in the case of the ErgoScope work simulator, it was a personal data recording, while the LVA allowed even the possibility of telephone inquiries).

The measurement results from different sources are compiled into a single standard data file so that the ATOM-CORE (Gergely & Takács, this special issue) can handle them in a suitable form.

Background information to the case studies

The organizations involved in the five case studies were the following:

1. KÉZMŰ, FŐKEFE, ERFO Plc. (in short: KÉZMŰ)
2. ATOMIX Fire and Damage Prevention Department Plc. (in short: ATOMIX)
3. MPT Postal Saving Security and Logistics Plc., within Budapest (in short: MPT1)
4. MPT Postal Saving Security and Logistics Plc., outside Budapest (in short: MPT2)
5. National Rehabilitation and Social Office (in short: NRSZH)

The applied measuring instruments with their short descriptions concerning the case studies are listed in *Table 2.*

Table 2. The applied measuring instruments in each case study

Measuring instruments	**Description**	**Case studies**				
		1.	**2.**	**3.**	**4.**	**5.**
Paper-pencil cube rotation task	A paper-and-pencil test is suitable for examining spatial manipulation ability (mental rotation) (Peters et al., 1995)	x				
MaxWhere cube rotation task (using a laptop)	A 3D-based cube rotation test is suitable for examining spatial manipulation ability (mental rotation) (MaxWhere, 2022)	x				
Social Network Analysis	A method for studying the dynamics, internal structures and other characteristics of different social networks (Czabán & Nagybányai Nagy, 2021)	x		x	x	
Anima questionnaire	General personality test	x				
BFI (Big Five Inventory)	A test for assessing the basic dimensions of personality (John & Srivastava, 1999)	x				
MET (Mental Health Test)	A test to assess psychological well-being and mental health (Vargha et al., 2020)	x				
RMMT questionnaire (Short Work Motivation Test)	Questionnaire for measuring work motivation	x				
Brengelmann questionnaire	A questionnaire suitable for measuring basic general personality traits (Brengelmann, 1959)		x			
Anger questionnaire	A test suitable for measuring the ways of expressing anger and rage		x			
Broadbent questionnaire	A scale suitable for measuring an individual's tendency to make cognitive mistakes		x			
Belbin questionnaire	A test for measuring behavior in work groups (Furnham et al., 1993)		x			
Eysenck questionnaire	A test suitable for measuring the two human supertraits (Extraversion and Emotional stability) and related dimensions (Eysenck & Eysenck, 1964)		x			
Type A-B personality questionnaire	A test for measuring type A and type B behavior		x			
Buss – Durkee hostility questionnaire	A questionnaire suitable for measuring the level of aggressiveness (Buss & Durkee, 1957)		x			
Maslach questionnaire	A suitable test for measuring the level of burnout (Maslach et al., 1997)		x			

Measuring instruments	**Description**	**Case studies**				
		1.	**2.**	**3.**	**4.**	**5.**
Assertiveness questionnaire	Questionnaire for measuring social efficiency		x			
Big Five (NEO-PI-R) questionnaire	A test suitable for measuring the five basic general, comprehensive personality traits (Costa & McCrae, 2008)		x			
D2 attention test	An attention test suitable for measuring the speed of information processing, rule- following and qualitative aspects of performance (Bates & Lemay, 2004)		x			
ÁSZVEK questionnaire	A questionnaire characterizing basic general personality traits measured using the General Personality Effectiveness and Leadership Virtues Questionnaire			x	x	
ErgoScope	Work simulator, work ability testing system, which examines the test subject in simulated situations (Izsó et al., 2015)			x	x	
LVA	Layered sound analysis technology, which can be used to determine the characteristics derived from sound segments that measure emotional and mental tension (Nemesysco, 2022)			x	x	
Communication Status Questionnaire	A questionnaire measuring the basic dimensions of human-to-human communication (Somlai, 2019)			x	x	
Conflictometer	The EM-05.58K (manufactured by STRUCTURE) desktop Complex sensorimotor tester and conflictometer (Burtaverde & Mihaila, 2011)			x	x	
RMSK questionnaire	Questionnaire for measuring the characteristics of occupational stress (Bilkei et al., 2000)			x	x	
Fixed interviews compiled by social experts						x

Source: edited by using own research data

Description of the samples

During the case studies, samples of different sizes were available to us. In these cases, both the sample size and its homogeneity along either application or other characteristics were significantly different. More detailed and accurate descriptions of these are available in the case studies themselves in Hungarian. The main characteristics of the samples, available to us in all case studies, are included in the table below.

Table 3. The studied jobs in each case study

1. KÉZMŰ		
Sample size	N	%
	120 persons	100
Studied job	box makers: 120 persons	100
2. ATOMIX		
Sample size	N	%
	74 persons	100
Studied job	fire fighters: 74 persons	100
3. MPT1 (within Budapest)		
Sample size	N	%
	215 persons	100
Studied jobs	value carriers: 92 persons	43
	value storage workers: 23 persons	11
	value managers: 42 persons	20
	money processors: 36 persons	19
	others: 22 persons	7
4. MPT2 (outside Budapest)		
Sample size	N	%
	202 persons	100
Studied jobs	value carriers: 131 persons	64
	value managers: 35 persons	17
	others: 36 persons	17
5. NRSZH		
Sample size	N	%
	16,431 disabled persons	100
Currently working	3,663 disabled persons	29
Never worked	348 disabled persons	3
More than fifteen years of employment	12,734 disabled persons	68

Note: N = absolute frequency; % = relative frequency
Source: edited by using own research data

Results

Frequency distributions of job success scales

As mentioned earlier, in the case of ATOM, competing algorithms are running (Gergely & Takács, this special issue), so the prediction and classification tables will provide an accurate comparison. The following classification results were obtained in the five samples included in the case studies.

Table 4. The frequency distributions of job success scores[1]

	Frequency distributions along the originally used 5-point job success scales						
Case study ↓	**1**	**2**	**3**	**4**	**5**	**Total**	*** Overall hit probability (for the derived two-point scales)**
1.	7	5	23	18	67	120	79.2%
2.	3	21	24	20	6	74	77.0%
3.	0	6	55	100	54	215	73.5%
4.	23	2	11	71	95	202	88.6%
5.	**Frequency distribution along the originally used 3-point job success scale**			**Frequency distribution along the derived 2-point job success scale**			
	1	**2**	**3**	**1**	**2**		*** Overall hit probability (for the derived two-point scale)**
	7,012	3,283	6,071	7,012	9,354	16,366	99.95%

Note:

* These data belong to the best-performing ("winner") ML algorithms. It can be seen that all values are much higher than the 50% chance probability and *"high enough"* (greater than 70%).

Source: edited by using own research data

These high overall hit probabilities, however, represent only the necessary condition for practical usability.

Even if *overall hit probabilities* are *"high enough",* as in this table, it could still happen that *sensitivity* or *specificity* is unacceptably low, as we will see later in the case studies.

[1] The frequency distributions of job success scores along the originally used 5-point and 3-point scales with the overall hit probabilities for the corresponding two-point scales in each case study. The originally used 5-point and 3-point scales were transformed into appropriate two-point scales. While all the 5-point job success scales were based on workplace leaders' judgments, the data on the 3-point motivation scale came from self-reporting.
The overall hit probabilities, corresponding to the default 50% cut-off level, are presented as percentages.

In such cases, selecting a better cut-off level on the *labelling probabilities* (and thus also producing a new related classification table) could be the solution depending on the particular prediction goals. As we can see later in *Table 9*, this method was working in the 1st and 2nd case studies but was not working in the 3rd (MPT1) and 4th (MPT2) case studies. It can be stated for these last two case studies that the sufficient condition for practical usability is not met.

Table 5. The main results at KÉZMŰ

With a cut-off p_1 = 0.5	**Case categorised by ATOM as in category 1**	**Case categorised by ATOM as in category 2**	**Total**
Case is actually in category 1	22	13	35
Case is actually in category 2	12	73	85
Total	34	86	120

With a cut-off p_1 = 0.225	**Case categorised by ATOM as in category 1.**	**Case categorised by ATOM as in category 2.**	**Total**
Case is actually in category 1	31	4	35
Case is actually in category 2	29	56	85
Total	60	60	120

Source: edited by using own research data

Results at KÉZMŰ (1)

It can be observed that the choice of the cut-off point here matters a lot. When should we consider someone a potentially "successful" or "unsuccessful" employee? At what actual probability do we call the expected performance acceptable?

In the upper part of the table (p_1 = 0.5), as it can easily be calculated, the *overall hit probability* is 95/120 = 0.792 (see also *Table 4*). Furthermore, ATOM can predict the failure (category 1) relatively badly (22/35 = 0.628), but the success (category 2) quite well (73/85 = 0.859). However, this company – since they have to employ almost every candidate for this job – was not interested in predicting success, but in predicting failure, (identifying those who should not be employed in any case, not even when the company is in strong need of workforce).

Selecting an appropriate cut-off point, and predicting failure can be radically improved. In the lower part of the table (p_1 = 0.225), the *overall hit probability* is only slightly lower (87/120 = 0.725), but the prediction of failure became much better (31/35 = 0.856).

Apart from these particular cut-off points, the *Overall Model Quality* was characterised by the *ROC* and the *Precision – Recall* curves as can be seen in *Figure 1* below.

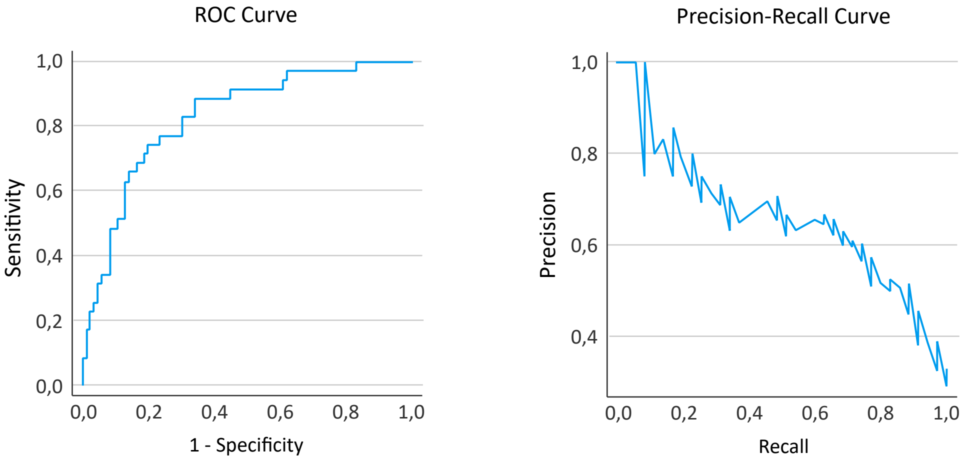


Figure 1. The *ROC* and the *Precision – Recall* curves for the KÉZMŰ study[2]

Here an example is presented for interpreting *ROC* curves and *Precision – Recall* curves in *Figure 1* at varying levels of cut-off points on their axes.

On the ROC curve it can be seen that if only a maximum 0,10 *false alarm probability* (1 – *specificity*) cut-off level can be accepted, the related *true alarm probability (sensitivity)* is maximally about 0.35. But if a maximum 0.20 *false alarm probability* can be tolerated, the related *true alarm probability* can grow up to about 0,68. Similarly, if a maximum 0.40 *false alarm probability* can be permitted, the related *true alarm probability* can be as high as about 0.90. Or, on the other way around, we can conclude that if we need at least about 0.35 *true alarm probability,* the price we have to pay for it is to accept at least 0.10 *false alarm probability,* etc.

On the *Precision – Recall* curve it can be seen that the precision is perfect (1.00) only below the 0.07 *recall (true alarm probability, sensitivity)* value. Similarly, to about a 0.40 *recall* value belongs about a 0.65 *precision.*

Results at ATOMIX (2)

Table 6. The main results at ATOMIX

With a cut-off of $p_1 = 0.5$	**The case categorised by ATOM as in category 1**	**The case categorised by ATOM as in category 2**	**Total**
The case is actually in category 1	57	0	57
The case is actually in category 2	17	0	17
Total	74	0	74

[2] Since the sample is only slightly imbalanced, even the ROC curve is interpretable. Both curves show acceptable prediction performance: the AUC (area under the curve) values – as the measure of *Overall Model Quality* – are high enough (for ROC: 0.827; for *Precision – Recall*: 0.750).

With a cut-off of $p_1 = 0.775$	**The case categorised by ATOM as in category 1**	**The case categorised by ATOM as in category 2**	**Total**
The case is actually in category 1	31	26	57
The case is actually in category 2	2	15	17
Total	33	41	74

Source: edited by using own research data

In the upper part of the table (p_1 = 05) the following can be seen: while the *overall hit probability* is 57/74 = 0770 (see also *Table 4*), ATOM categorised all cases as being in category 1. It means that under the given circumstances the model cannot differentiate between the two categories. Since this company was interested in predicting the job success of candidates as accurately as possible, to meet this requirement we had to find another cut-off point.

As can be seen in the lower part of the table, selecting p_1 = 0775 is a good solution to this problem. In this case, the prediction of job success becomes quite high: 15/17 = 0882.

Apart from these particular cut-off points, the *Overall Model Quality* was characterised by the *ROC* and the *Precision – Recall* curves as can be seen in *Figure 2* below.

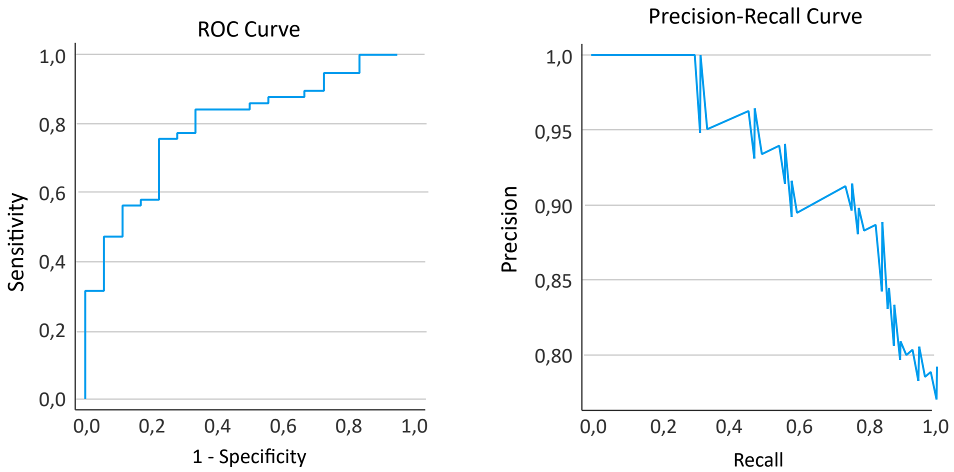


Figure 2. The *ROC* and the *Precision – Recall* curves for the ATOMIX study[3]

[3] Since the sample is only slightly imbalanced, even the ROC curve is interpretable. Both curves show acceptable prediction performance: the AUC (area under the curve) values – as the measure of *Overall Model Quality* – are high enough (for ROC: 0.787; for *Precision – Recall*: 0.670).

Results at MPT (3, 4)

In the case of the MPT company two separate ATOM studies were conducted: the first involved 215 employees within Budapest (MPT1), while the second involved 202 employees outside Budapest (MPT2).

Table 7. The main results at MPT (all these data are based on the default 50% cut-off level)

MPT1 (within Budapest)	The case categorised by ATOM as in category 1	The case categorised by ATOM as in category 2	Total
The case is actually in category 1	153	1	154
The case is actually in category 2	56	5	61
Total	209	6	215

MPT2 (outside Budapest)	The case categorised by ATOM as in category 1	The case categorised by ATOM as in category 2	Total
The case is actually in category 1	179	0	179
The case is actually in category 2	21	0	23
Total	202	0	202

Source: edited by using own research data

However, the evaluation policy can also influence the algorithm's behaviour. For example, in the case of the MPT, there are very mixed jobs, so the criteria for the actual salary differ significantly in the different jobs. In such cases, the practice is not necessary to filter out the best, most excellent employees in the system, but those whom we do not want to employ for some reason.

Although choosing cut-off levels, other than the default 50%, resulted in slightly improved results, taken overall, these results are still unacceptable.

Both MPT1 and MPT2 samples were somewhat imbalanced. Therefore, ROC curves were not considered. The *Precision-Recall* curves were created instead, but these showed for MPT1 weak-medium (AUC: 068) and for MPT2 unacceptably low (AUC: 041) prediction performance. Because of all these deficiencies, the related graphs are not presented.

The reasons for these inadequate, and partly useless models are very probably that both the job success data and the predictors were of rather low quality:

1. the job success data, because the leaders who gave the scaled judgments, unfortunately, had different criteria for rating;
2. the predictors, because we found signs of random answers by many employees to test questions.

Results at NRSZH (5)

The aim of this study was – as indicated in (Izsó, this special issue) – not to predict job success, but to predict work motivation (intention to return to work). Here is the meaning of the categories: 1 = *"less likely to return to work"*, 2 = *"more likely to return to work"*.

Table 8. The main results at NRSZH

With a cut-off of $p_1 = 0.5$	The case categorised by ATOM as in category 1	The case categorised by ATOM as in category 2	Total
The case is actually in category 1	7,012	0	7,012
The case is actually in category 2	8	9,346	9,354
Total	7,020	9,346	16,366

Source: edited by using own research data

Here we were able to query and test the data of 16,366 persons, and it is clear from the results that we do not need to carry out any further testing here. Overall, the ATOM's model worked extremely well, there were only 8 persons – out of the 16,366 (!) – who were incorrectly identified.

The reasons for these almost perfect results were (1) the relatively homogeneous sample (all involved persons were disabled), (2) the high-quality predictors (collected by highly experienced social workers) and (3) the very large sample size.

As can be seen in *Figure 3* below, the *ROC* and the *Precision – Recall* curves show quite exceptionally good, practically perfect, *Overall Model Quality.*

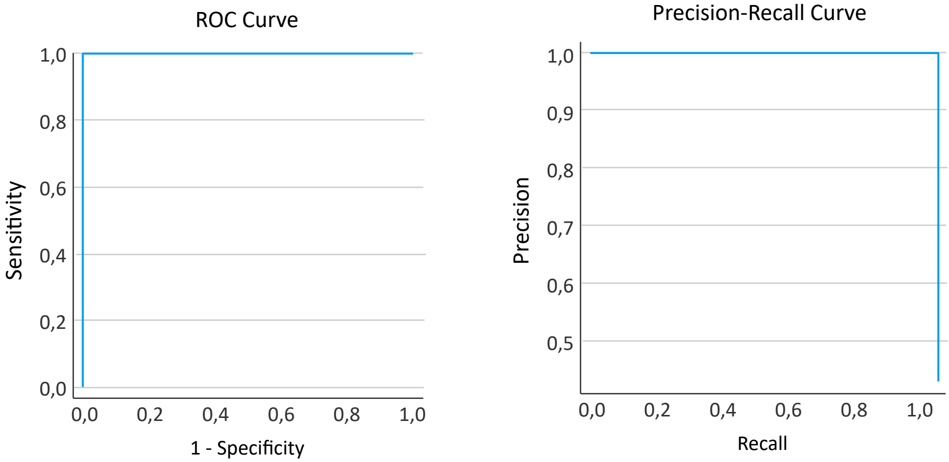


Figure 3. The *ROC* and the *Precision – Recall* curves for the NRSZH study[4]

It has to be mentioned, that earlier we have done some research works with entirely different goals based on this same database. The results of this research of different focus were also published: Pósfai et al. (2013), Kertész et al. (2017).

[4] Due to the high quality predictors and the extremely large sample size, both the *ROC* and the *Precision – Recall* curves show practically perfect prediction performance.

Table 9. The summaries of the main results of the five case studies

Classification tables for two-point scales integrated into one complex table for all case studies							
	1. KÉZMŰ						
	Cut-off point p_1 = 0.5				Cut-off point p_1 = 0.225		
	1	2	Total		1	2	Total
Actual 1	22	13	35	Actual 1	31	4	35
Actual 2	12	73	85	Actual 2	29	56	85
Total	34	86	120	Total	60	60	120
Goal: to improve failure prediction probability	Failure prediction probability = 0.628				Failure prediction probability = 0.856		
	2. ATOMIX						
	Cut-off point p_1 = 0.5				Cut-off point p_1 = 0.775		
	1		Total		1	2	Total
Actual 1	57		57	Actual 1	31	26	57
Actual 2	17		17	Actual 2	2	15	17
Total	74		74	Total	33	41	74
Goal: to improve success prediction probability	Success prediction probability = 0.000				Success prediction probability = 0.882		
	3. MPT1 and 4. MPT2						
	Cut-off point of p_1 = 0.5				Cut-off point p_1 = 0.5		
	1	2	Total		2		Total
Actual 1	5	56	61	Actual 1	23		23
Actual 2	1	153	154	Actual 2	179		179
Total	6	209	215	Total	202		202
Goal: to improve failure prediction probability	Improvement was not possible by changing cut-off point						
	5. NRSZH						
	Cut-off point p_1 = 0.5						
		1		2		Total	
Actual 1		7,012		0		7,012	
Actual 2		8		9,346		9,354	
Total		7,020		9,346		16,366	
Goal: to provide accurate prediction for both category		No need for improvement (already almost perfect)					

Source: edited by using own research data

Summaries of main results

In summary, it turned out, that in the cases of KÉZMŰ and ATOMIX by selecting other suitable labelling probability cut-off points, instead of the default 50%, we were able to solve the problem quite well.

In the cases of MPT1 and MPT2, however, choosing other cut-off levels resulted only in slightly improved results, while the measure of *Overall Model Quality* (AUC of the *Precision – Recall* curves) for the MPT1 indicated a weak-medium, for the MPT2 an unacceptably low performance. Because of these deficiencies, the related results were omitted.

Finally, in the case of NRSZH, there was no need to change the cut-off level, the results were directly usable and interpretable. ATOM was able to build up an extremely effective model.

Discussion

In this section first (1) the main lessons learnt from the five real-life workforce selection case studies are discussed, and later (2) the limitations and possibilities of practical usability are summarised. Finally, (3) ATOM's perspectives in field applications and further development are outlined.

1. Most important lessons learnt from the five real-life case studies:
   a) *ATOM, as a prediction system,* is very susceptible to data quality. By this, we mean that:
      - Regarding jobs, their work content should be as homogeneous as possible. Heterogeneous analyses are like working with thoroughly mixed distributions, identifying them is not necessarily easy, and the content behind the intention may mean something else.
      - Regarding job success data, it is also worth making job success evaluations by the management as objective as possible. A Likert scale evaluation means something different, such as performance based on a quota and its band classification (compare, for example, the question How much do you value a good workforce? with the evaluation of the "grade received based on the percentage of graduation results").

   b) *The sample size* can decisively change some procedures' operation – thus also its predictive efficiency. This also supports our idea of working with competing algorithms (Gergely & Takács, this special issue) during evaluations. Our proposal for a sample size of about a minimum of 100, as a nice round number, of course, is not the result of some exact derivation. It is merely an experience-based approximate rule of thumb that is only valid if both the predictors and job success measures are of acceptable quality. We saw that for the 2nd case study (ATOMIX) a sample of only 74 firefighters was enough for ATOM to provide well-established useful practical results, because of the quite outstanding data qualities. On the other side, however, for the 3rd (MPT1) and 4th (MPT2) case studies, ATOM using samples of even 215 and 202, could not produce practically usable results. The probable reason for that was that both the job success data and the predictors were of rather low quality.

c) *The free choice of labelling probability cut-off points* showed significantly different decision patterns. That is why we decided not to provide the classification tables as information for employers (Gergely & Takács, this special issue; Pusker, Gergely & Takács, this special issue), but rather the success category (labelling) probabilities.

It was observed that both the strategy (looking for the best or the minimum entry-level) and the characteristics of the sample (the "success" category can be moved down or up) decisively determine the selection of the cut-off points.

The case studies demonstrated what an automated system, with well-defined performance indicators and honest responses from managers and employees, is capable of.

2. The limitations and possibilities of ATOM's practical usability:

It turned out clearly, that the main limitation concerning ATOM's practical usability is the requirement of a relatively large sample size (minimally about 100) for the ML algorithms for effective learning, and data quality.

These limitations can be quite restrictive. Only in a smaller part of all existing jobs work at least about 100 employees, whose work content is "homogeneous" enough (whose task and work activity is largely the same). The requirement of data quality is also often difficult to meet. Even using the simplest job success measures, the workplace leaders' subjective judgments, extra care must be taken to ensure the necessary reliability and validity. If objective performance data are used, the difficulties are not smaller, only different by nature.

We are facing similar challenges concerning the predictors. Again, considering the simplest predictors, scores of certain personality (or other) tests, we have to ensure reliable data collection (to prevent random answers and other biases, etc.).

If objective performance data are used as predictors, their relevance must be carefully checked. A good example of that is what we presented in (Izsó, Berényi & Pusker, this special issue): selecting appropriate objective performance parameters measured with the help of the ErgoScope work simulator (e.g. static and dynamic force measurements, grip strength, keyboard operation, turning/switching and button pressing, work capacity, monotony tolerance, etc.) can produce a more accurate prediction of job success by ATOM.

It can also be a limitation, that – by the applied business model – not the ATOM package itself, only its service is for sale. However, the ATOM's operational principle of applying multiple ML algorithms running in parallel and selecting the "winner", provides such flexibility that very probably represents a significant competitive advantage.

3. ATOM's perspectives in field applications and further development:

Of the *employees'*, the *employers'*, and the *experts'* functionalities of ATOM, in this special issue the *employees'* was not targeted at all. Although the employees' web-based data collection and feedback system – as a working prototype – is ready for larger-scale testing in the field of recruitment, up to now we have not had the possibility to carry out such systematical testing. One of our most urgent future tasks is just to complete these functional and usability testing, and later – based on the results of testing –, to further develop these services. This is partly true for the *employers'* decision functionalities, which still have to broaden (e.g., by installing appropriate new ML algorithms for further increasing flexibility, involving additional procedures, introducing new functions supporting longitudinal data analysis, etc.).

Concerning the *experts'* functionalities, since ATOM is basically designed for automatic prediction and not for explanatory purposes, our philosophy is not to build in newer and more sophisticated analysis tools. When such tools are needed in practice – very probably not too often – for additional analyses, we propose to use external statistical packages, like IBM SPSS Statistics, SAS, JASP, JAMOVI, R, PYTHON, etc. (as in this article we used IBM SPSS Statistics). The fact that ATOM identifies the best-performing "winner" ML algorithm, can provide a useful starting point for such additional analyses.

Összefoglalás

Szemléltető esettanulmányok az ATOM valós alkalmazására

Háttér és célkitűzések: Annak valós esettanulmányok útján történő bemutatása, hogy hogyan használható a gyakorlatban az ATOM a munkaerő kiválasztására.
Módszer: Az ATOM osztályozási (klasszifikációs) teljesítményének mérésére alkalmas metrikák meghatározása után – az egyszerűség, a lehető legnagyobb megbízhatóság és az egységesség érdekében minden esetben bináris beválási skálákat használva – öt konkrét, valós terepvizsgálat beválás-előrejelzési eredményeit mutatjuk be, elsősorban táblázatos formában.
Következtetések: A következő főbb gyakorlati tapasztalatok voltak megállapíthatók: (1) az ATOM érzékeny a felhasznált adatok minőségére, ezért minden szempontból megfelelő beválási kritériumokat és prediktorokat kell alkalmazni; (2) a tanító mintának legalább 100 személy megfelelő adataiból kell állnia; (3) a legjobb megoldások megtalálásának az a leghatékonyabb módszere, ha az egyes beválási kategóriákba történő illeszkedés valószínűségének skáláján mindig az adott problémának megfelelő vágási szinteket alkalmazzuk.
Kulcsszavak: ATOM, toborzás, munkaerő-kiválasztás, vágási szintek (cut-off levels), klasszifikációs teljesítmény

References of this Special Issue

Izsó, L. (2023). The concept of an AI-based expert system (ATOM) for predicting job success. *Alkalmazott Pszichológia*, *25*(3), 5–13.

Gergely, B. & Takács, Sz. (2023). ATOM – a flexible multi-method machine learning framework for predicting occupational success. *Alkalmazott Pszichológia*, *25*(3), 15–30.

Pusker, M., Gergely, B., & Takács, Sz. (2023). ATOM's structure – employee and employer feedback, survey site. *Alkalmazott Pszichológia*, *25*(3), 53–72.

Izsó, L., Berényi, B., & Pusker, M. (2023). Jointly applying a work simulator and ATOM to prevent occupational accidents and MSD through workforce selection. *Alkalmazott Pszichológia*, *25*(3), 73–91.

References

Alwin, D. F., Baumgartner, E. M., & Beattie, B. A. (2018). Number of response categories and reliability in attitude measurement. *Journal of Survey Statistics and Methodology*, *6*(2), 212–239.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*(1). 9–30.

Bates, M., & Lemay, E. P. (2004). The d2 Test of Attention: Construct validity and extensions in scoring techniques. *Journal of the International Neuropsychological Society, 10(3).* 392–400.

Bilkei P., Szabó B., & Böröcz I. (2000). *Rendvédelmi szervek munkahelyi stressz kérdőíve.* Útmutató *az indexek* értékeléséhez. Manuscript.

Brengelmann, J. C. (1959). Differences in questionnaire responses between English and German nationals. *Acta Psychologica, 16.* 339–355.

Burtaverde, V., & Mihaila, T. (2011). Significant differences between introvert and extrovert people's simple reaction time in conflict situations. *Romanian Journal of Experimental Applied Psychology, 2*(3). 18–25.

Buss, A. H., & Durkee, A. (1957). An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology, 21*(4). 343–349.

Costa, P., & McCrae, R. (2008). The revised NEO personality inventory (NEO-PI-R). In *The SAGE Handbook of Personality Theory and Assessment, 2.* (pp. 179–198).

Czabán Cs., & Nagybányai Nagy O. (2021). A pszichológiai szakirodalmi feldolgozás támogatása a hálózatelemzés módszerével: A tanácsadás pszichológiájának lehetséges taxonómiája. *Magyar Pszichológiai Szemle*, *76*(3–4). 549–567.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning.* Other related links: https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/ https://www.r-bloggers.com/2019/03/what-it-the-interpretation-of-the-diagonal-for-a-roc-curve/ https://towardsdatascience.com/on-roc-and-precision-recall-curves-c23e9b63820c

Eysenck, H. J., & Eysenck, S. B. G. (1964). *Manual of the Eysenck personality inventory.* University of London Press.

Eubanks, B. (2022). *Artificial Intelligence for HR. Use AI to support and develop a successful workforce*. Second Edition. Kogan Page Limited.

Furnham, A., Steele, H., & Pendleton, D. (1993). A psychometric assessment of the Belbin Team-Role Self-Perception Inventory. *Journal of Occupational and Organizational Psychology*, *66*(3). 245–257.

Henle, C. A., Dineen, B. R., & Dulffy, M. K. (2019). Assessing intentional resume deception: Development and nomological network of a resume fraud measure. *Journal of Business and Psychology*, *34*(1). 87–106.

Izsó, L., Székely, I., & Dános, L. (2015). Possibilities of the ErgoScope high fidelity work simulator in skill assessment, skill development and vocational aptitude tests of physically disabled persons (*„Best Paper Award”* winner conference paper). 13th International Conference of the Association for the Advancement of Assistive Technology in Europe, Sept. 9–12, Budapest, Hungary. As book chapter In Sik-Lányi, C., Hoogerwerf, E. J., Miesenberger, K., Cudd, P. (Ed.s), *Assistive Technology* (pp. 825–831). IOS Press.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In: L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research, Vol. 2*. (pp. 102–138). Guilford Press.

Kertész, A., Séllei, B., & Izsó, L. (2017). Key Factors of Disabled People's Working Motivation: An Empirical Study Based on a Hungarian Sample. *Periodica Polytechnica, Social and Management Sciences*, *25*(2). 108–116.

Maslach, C., Jackson, S., & Leiter, M. (1997). *The Maslach Burnout Inventory Manual*. The Scarecrow Press.

MaxWhere (2022, December 11). Virtual spaces with the benefits of reality. https://www.maxwhere.com/

Nemesysco (2022, October 23). Nemesysco voice analyst technologies. http://nemesysco.com/

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse Mental Rotations Test – different versions and factors that affect performance. *Brain and Cognition, 28*(1). 39–58.

Pósfai G., Séllei B., & Kertész A. (2013). A megváltozott munkaképességű emberek munkamotivációját befolyásoló kognitív és érzelmi tényezők. *Alkalmazott Pszichológia*, *15(4)*. 47–57.

Somlai R. (2019). Vezetői stílusok vizsgálata személyes készségek elemzésével. *Taylor Gazdálkodás-* és *Szervezéstudományi Folyóirat*, *1*(35). 85–96.

Tasdemir, F. (2015). A Study for Developing a Success Test: Examination of Validity and Classification Accuracy by ROC Analysis. *Procedia – Social and Behavioral Sciences, 191*. 110–114.

Vargha A., Zábó V., Török R., & Oláh A. (2020). A jóllét és a mentális egészség mérése: a Mentális Egészség Teszt. *Mentálhigiéné* és *Pszichoszomatika*, *21*(3). 281–322.

BLANKA BERÉNYI

BENCE GERGELY

LAJOS IZSÓ

JUDIT T. KÁRÁSZ

MÁTÉ PUSKER

SZABOLCS TAKÁCS

AUTHORS