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.

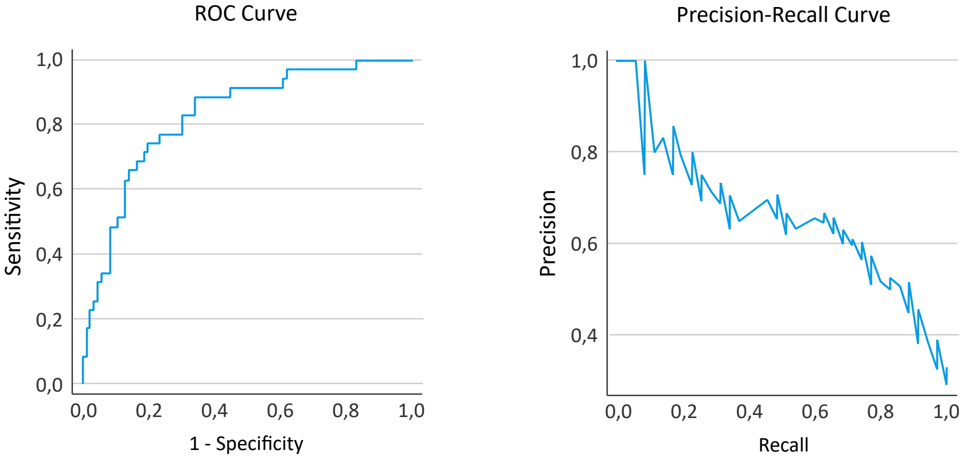

Figure 1. The *ROC* and the *Precision – Recall* curves for the KÉZMŰ study[2]

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.

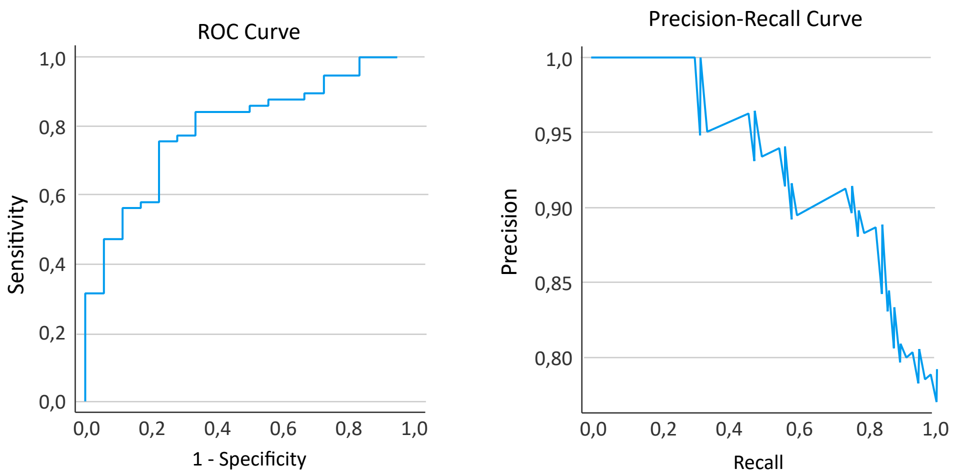

Figure 2. The *ROC* and the *Precision – Recall* curves for the ATOMIX study[3]

[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.*

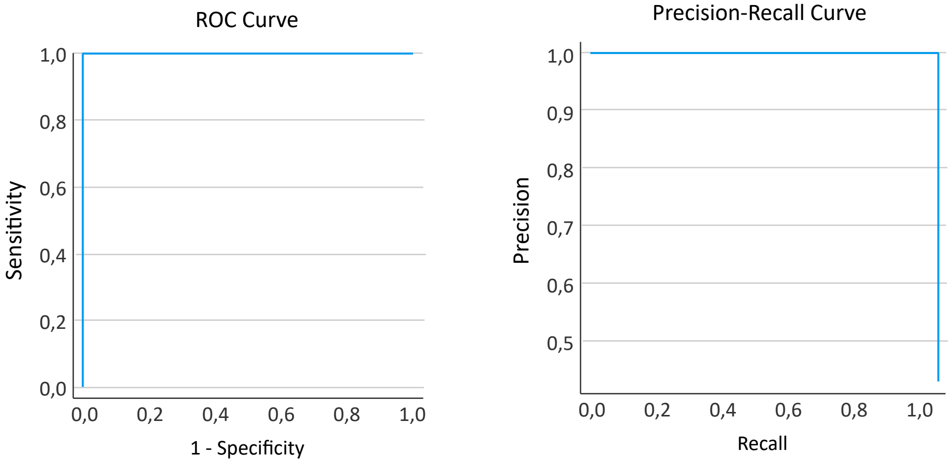

Figure 3. The *ROC* and the *Precision – Recall* curves for the NRSZH study[4]

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.