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			**Math Proficiency Level, CLOSED**								
1st			**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**			**Total**
Math Profici-ency 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			**Reading Proficiency Level, CLOSED**								**Total**
1st			**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**			
Reading Proficiency 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			**Math Proficiency Level, CLOSED**								
1st			**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**			**Total**
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			**Math Proficiency Level, CLOSED**								
1st			**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**			**Total**
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