Wojciech Malec

Katolicki Uniwersytet Lubelski Jana Pawła II

ASSESSMENT IN ONLINE EDUCATION: FOCUS ON SCORE REPORTING

OCENIANIE W NAUCZANIU NA ODLEGŁOŚĆ: KWESTIA RAPORTOWANIA WYNIKÓW

This article discusses score reporting in the context of computer-based assessment, especially online language testing in academic environments. Score reporting, understood as the communication of information about test results, is presented as an essential element of operational test use, which is a component of the process of test development. The article gives some consideration to the calculation of two types of final results, i.e. arithmetic and weighted mean scores, and outlines an online testing system, WebClass, where such results are obtained and reported automatically. Moreover, examples are given of several score reports, such as individual single-test reports and cumulative reports (summarising the results of all assessments) for students, as well as cumulative group reports and reports summarising the results of a selected test for teachers. Thanks to technology, score reports can be automatically updated as soon as new responses are submitted by test takers.

Keywords: online language assessment, test development, score reporting

1. Introduction

The last two years have seen an increased level of interest in e-education, presumably due in large measure to the transition from onsite to online classes brought about by the COVID-19 pandemic. This interest is evidenced by numerous publications dealing with the use of technology for language learning and teaching (e.g. Fuchs, Hauck, Dooly, 2021; Lütge, Merse, 2021; Sendur, Kościńska, 2021), many which focus specifically on various aspects of language testing and assessment in digital environments (e.g. Isbell, Kremmel, 2020; Soproni, 2020; Bahari, 2021; García Laborda, Fernández Álvarez, 2021). Admittedly, the sudden switch to web-based assessment that recently occurred in formal teaching contexts in many countries, including Poland, was not made at the discretion of teachers and learners, but rather resulted from pandemic restrictions imposed on educational institutions. On the other hand, it is worth pointing out that computer-based testing (CBT) in general and web-based testing (WBT) in particular have long been recognized as having some clear advantages over paper-based testing (PBT). Accordingly, the choice of CBT does not need to be a necessary evil (Malec, 2020).

An oft-cited advantage of CBT is the provision of accurate score reporting (Brown, 2016), which includes, but is not limited to, "a summary of the points scored" (Romaniuk, Łukasiewicz-Wieleba, 2021, p. 26). However, somewhat surprisingly, academic teachers surveyed by Klimowicz (2020) failed to mention score reporting as a benefit of online assessment. This may be because of the fact that the assessment tools which are commonly used for testing do not allow teachers to take full advantage of the possibilities afforded by technology. For example, online tests and quizzes are often created and administered using Microsoft Forms or Google Forms (Kita, 2021), yet the usefulness of score reports provided by tools of this kind is fairly limited, unless we only need basic information about the number of points scored on each individual assessment. An arguably better solution is to use a learning management system (LMS), such as Google Classroom or Moodle, which offers gradebooks allowing educators to record and manage students' grades. This article discusses another LMS, namely WebClass (webclass.co), where individual and group score reports can be generated for the assessments administered in a period of study.

2. Score reporting as a (sub)component of test development

Test development can be conceptualized as a cyclical process consisting of four major components, each of which is associated with a range of activities (Malec, 2018; see also Lane, Raymond, Haladyna, Downing, 2016). These components include design (making plans for the test and drawing up a list of specifications), production (writing tasks or items and compiling the test), operational use (using the test for its intended purposes), and evaluation (assessing the quality of the test and individual items). Reporting test scores is part of the third component, along with test administration, scoring, and decision making. It can thus be viewed as a subcomponent of operational test use.

In summative assessment, score reporting and decision making are closely linked because test scores provide the basis for making inferences and decisions about the test takers. This means, to some extent at least, that by reporting test results we also report test-based decisions. In classroom settings, whether online or onsite, a variety of classification decisions can be made, the most basic of which are 'master' (pass) and 'non-master' (fail). It should be added that those decisions are rarely based on raw scores. Instead, the points obtained are often converted into percentages, percentiles, or standardized scores. Moreover, using a reliability or dependability estimate, the standard error of measurement (*SEM*) can be calculated, either for relative or for absolute decisions (e.g. Bachman, 2004). The value of the *SEM* can then be used to obtain a confidence interval (*CI*), i.e. a score range within which the test taker's true score or domain score is likely to fall with a specific degree of probability. In classroom contexts, the *SEM* and *CI* may be useful when we need to make decisions on the basis of a specific cut-score (e.g. the pass/fail boundary). For example, some test results below a given pass mark may still be within a specific *CI*, and therefore non-mastery decisions based solely on such results will not be fully justified. In fact, Brown and Hudson (2002) recommend that borderline test takers be given additional assessments in order to obtain further evidence in support of the classification decisions.

All the required calculations as well as decisions regarding score interpretations are followed by reports of the results. These are of various types, depending on the situation (Carr, 2011). For example, reports may be intended either for individuals or for groups of students. They may be either digital or paper-based. A variety of information can be included in a score report (Zenisky, Hambleton, 2016; see also Deng, Yoo, 2009, for a bibliography of resources for reporting test scores). Generally speaking, reports may include both descriptive elements and performance indicators. Descriptive elements contain information about the test (such as name,

date, purpose, and the like) as well as details identifying the group or individual for which the report is intended. Performance indicators, on the other hand, may include summative results, grades, pass-fail decisions (along with a cut-score), performance-level descriptions, as well as formative or diagnostic information. Summative results are usually test scores, the basic types of which are raw and transformed scores. The latter include weighted scores, percentages, percentiles, stanines, as well as standardized scores such as z-scores or T-scores (more on these can be found in, e.g., Brown, 1996).

2.1. Calculating final results

Reports compiled at the end of a semester or some other period of study typically give information about the scores obtained on all the assessments administered as well as a final result in the form of an average score (which provides the basis for the final grade).[1] In school contexts, two types of average scores are commonly used, namely arithmetic and weighted means.

The arithmetic mean is calculated by adding up all the scores (placed on the same scale, e.g. percentage correct) and dividing the result by the number of assessments. This type of mean assumes that every assessment contributes equally to the final result. The use of the weighted mean, by contrast, is predicated on the assumption that certain tests and assessments are worth more than others and should be given priority. The particular weights (in the form of multipliers or percentages) can be pre-established by test developers, and regarded as fixed and unrelated to test length (the number of points to score). However, it is also possible to implement a different procedure, one which assumes that the weights are potentially variable and dependent on the maximum possible scores. More precisely, the weights can be computed for each assessment by working out the highest possible aggregate score (i.e. maximum score on all the assessments taken together) and calculating the proportion (or percentage) that each particular maximum contributes to the aggregate maximum. With the help of technology, the whole procedure can be automated and a score reporting system can easily adjust the weights every time a new assessment is added to the dataset. By way of illustration, Table 1 presents examples of three tests for which the weights have been obtained using the procedure in question. Each particular weight is a reflection of the test's importance. As can be seen in Table 1, the weights correlate with the maximum possible scores on the tests.

[1] Similarly, when a single test or exam consists of several components (e.g. grammar, vocabulary, reading), the final result is often a mean of all the individual component scores.

Table 1. Test weighting determined on the basis of the maximum possible score

Assessment	Max	Aggregate Max	Weight
Test 1	10	60	17%
Test 2	20		33%
Test 3	30		50%

The next step is to obtain the weighted mean score for a particular student, i.e. the final result of all the assessments taken together. The calculation involves multiplying the student's test scores by the respective weights and then adding up the products. It is worthy of note that the weighted-mean approach to calculating the final result is, in effect, like treating the individual tests as subtests or components of one larger test.

Furthermore, the weighted mean pass mark, which may be required for the final grade, can be computed in a similar manner. For example, suppose we have the following two tests:

Table 2. Data for the calculation of the weighted mean pass mark (Set 1)

	Max	Pass Mark	Weight
Test 1	20	14 (70%)	25%
Test 2	60	18 (30%)	75%

The weights for the tests in Table 2 have been obtained using the procedure illustrated in Table 1. Now, if we multiply each percentage pass mark by a factor representing the weight of the respective test, and then sum the products, we get the following result:

$$70\% \times 0.25 + 30\% \times 0.75 = 17.5\% + 22.5\% = 40\%$$

Thus, the weighted mean pass mark for these two tests taken together is 40%. As mentioned above, this approach is tantamount to treating the two tests as components of one larger test. Therefore, we can also obtain the overall percentage pass mark by summing the individual cut-scores (expressed in points), dividing the result by the aggregate score, and then multiplying by 100:

$$\frac{14 + 18}{20 + 60} \times 100 = \frac{32}{80} \times 100 = 0.40 \times 100 = 40\%$$

Interestingly, if we swap the two tests' pass marks, we obtain a completely different weighted mean pass mark. Consider the data in Table 3:

Table 3. Data for the calculation of the weighted mean pass mark (Set 2)

	Max	Pass Mark	Weight
Test 1	20	6 (30%)	25%
Test 2	60	42 (70%)	75%

Using the second procedure given above, we can obtain the weighted mean pass mark for the data in Table 3:

$$\frac{6\ +\ 42}{20\ +6\ 0} \times 100 = \frac{48}{80} \times 1\ 00 = 0\ .60 \times\ 100 = 60\%$$

Clearly, the larger test (with more points to score) has a greater impact on the weighted mean pass mark. By way of comparison, the arithmetic mean of the pass marks on the two tests in Table 2 and Table 3 would in both cases be the same:

$$(70\% + 30\%) \div\ 2 =\ 50\%$$

It should be pointed out that if we treat the two tests as independent measures of equal importance, we should take the arithmetic-mean approach.

Some of the above calculations would be slightly more straightforward if all of the assessments had the same pass mark. In such a case, there would be no difference between the arithmetic and weighted means of the individual pass marks. Similarly, if each assessment had the same number of questions (and the same maximum number of points to score), there would be no difference between the arithmetic and weighted means of the total test scores.

3. Reporting test results on WebClass

This section looks at score reports on WebClass. The reports can be produced either for each single test or cumulatively for all the tests administered in a specific time period, which is typically one semester. Individual reports can be accessed by students and teachers. For teachers, additionally, the system generates reports for entire groups of learners as well as for tests, either with a summary of students' scores or with the outcome of quantitative analysis (for test and item evaluation). The reports are automatically updated every time new responses are submitted.

3.1. WebClass

WebClass can be categorized as an academic LMS (cf. Foreman, 2018) incorporating features such as learner management, communication, content authoring, and assessment. It has recently been utilised primarily as an online assessment tool to develop web-based language tests for university and secondary school learners (Malec, Malec, 2021). The testing component of the platform can be used to create tests and quizzes of various types (including selected- and constructed-response items as well as extended-response tasks), administer them to groups of students, monitor the test-taking process, mark the responses automatically (but with a possibility of verifying and overriding the scoring), provide feedback to the test takers, analyse tests and items statistically, store items in item pools and then reuse them with the aid of a test generator or by importing them into existing tests (further details can be found in Malec, 2018). The next two sections expand on how the system reports results to students and teachers.

3.2. Reports for students

Students can access individual single-test reports after submitting their responses, assuming that the teacher or test administrator has activated the relevant setting in the test manager. With multiple submissions of the same test, a separate report is available for each set of responses. The reports begin with basic information about the test and points scored, followed by the test itself, i.e. the test items, the student's responses, automatic corrections displayed in speech bubbles (in the case of limited-production items, e.g. gap-filling, scored against a key), additional formative information in the form of teacher's comments provided as general and answer-specific feedback, as well as item scores. The entire report is similar to a traditional paper-based test marked by a teacher (see Figure 1).

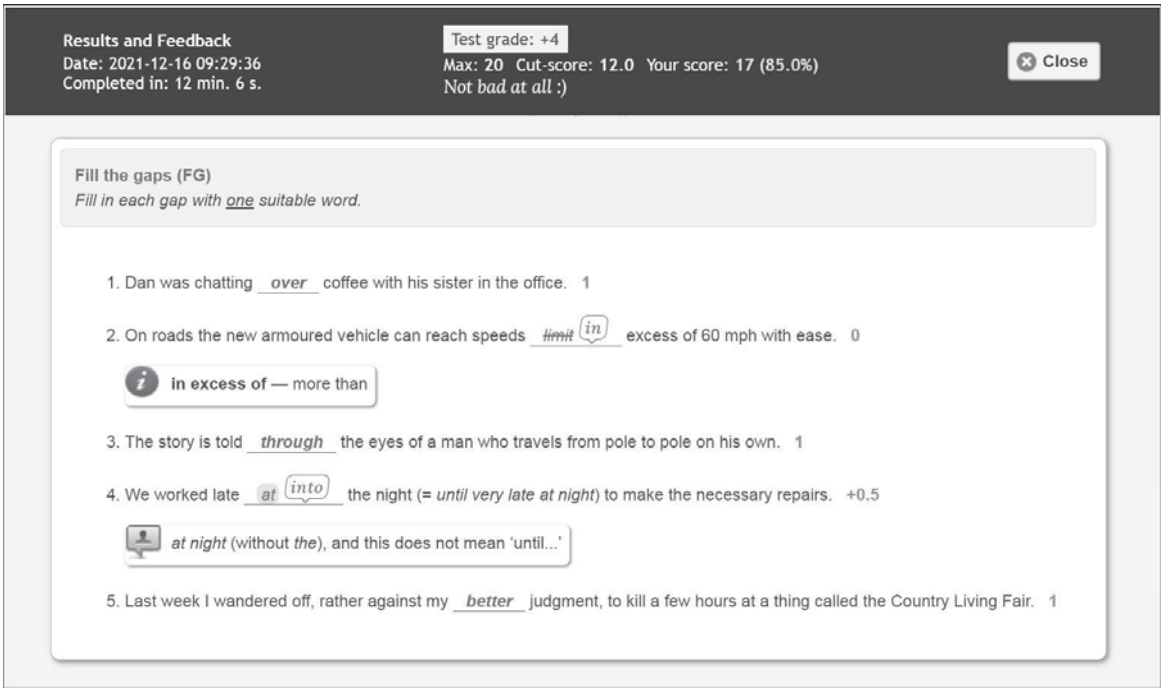


Figure 1. Part of an individual single-test report

Individual cumulative reports, on the other hand, summarize the results of all the assessments hitherto administered. In addition, they include information about the number of classes (or lessons) skipped and unexcused.[2] Moreover, the reports show the outcome of self-assessment (see also Malec, 2015, on how students can use the platform to self-evaluate their achievement of a set of learning objectives). An example of such a report is given in Figure 2.

[2] Teachers can access an attendance grid available for each registered group or class to keep track of the learners' absences. The number of classes skipped is then updated automatically in the reports.

Classes skipped (unexcused): **1**

Here are the results of your tests, quizzes, etc:

Include practice: ☐

Assessment	Max	Pass	Score	Grade
Attendance	20	12.0	22.0 (110.0%)	☺
English at Work (MC)	5	3.0	5.0 (100.0%)	5
English at Work (FG)	5	3.0	5.0 (100.0%)	5
Kahoot! (Unit 1)	10	6.0	12.0 (120.0%)	☺
Unit 1	30	18.0	27.5 (91.7%)	+4
Big Test 1	50	30.0	44.5 (89.0%)	+4
Unit 2	20	12.0	15.0 (75.0%)	+3
Presentation	20	12.0	18.8 (94.2%)	5
Unit 3	25	15.0	24.0 (96.0%)	5
Big Test 2	50	30.0	45.0 (90.0%)	+4
Final	100	60.0	93.0 (93.0%)	5
		Arithmetic mean:	**96.3%**	**5**
		✔ **Weighted mean:**	**93.1%**	**5**

NOTE:
The smiling face means that bonus points have been awarded.

Component	Weight	Mastery level		Performance result		Final pass	Final result	Grade
		Overall	Component	Overall	Component			
Teacher-administered assessments	95%	60.0%	57.0%	93.1%	88.4%	60.0%	92.0%	**5**
Self-assessment	5%	60.0%	3.0%	71.3%	3.6%			

Figure 2. An individual cumulative report on WebClass

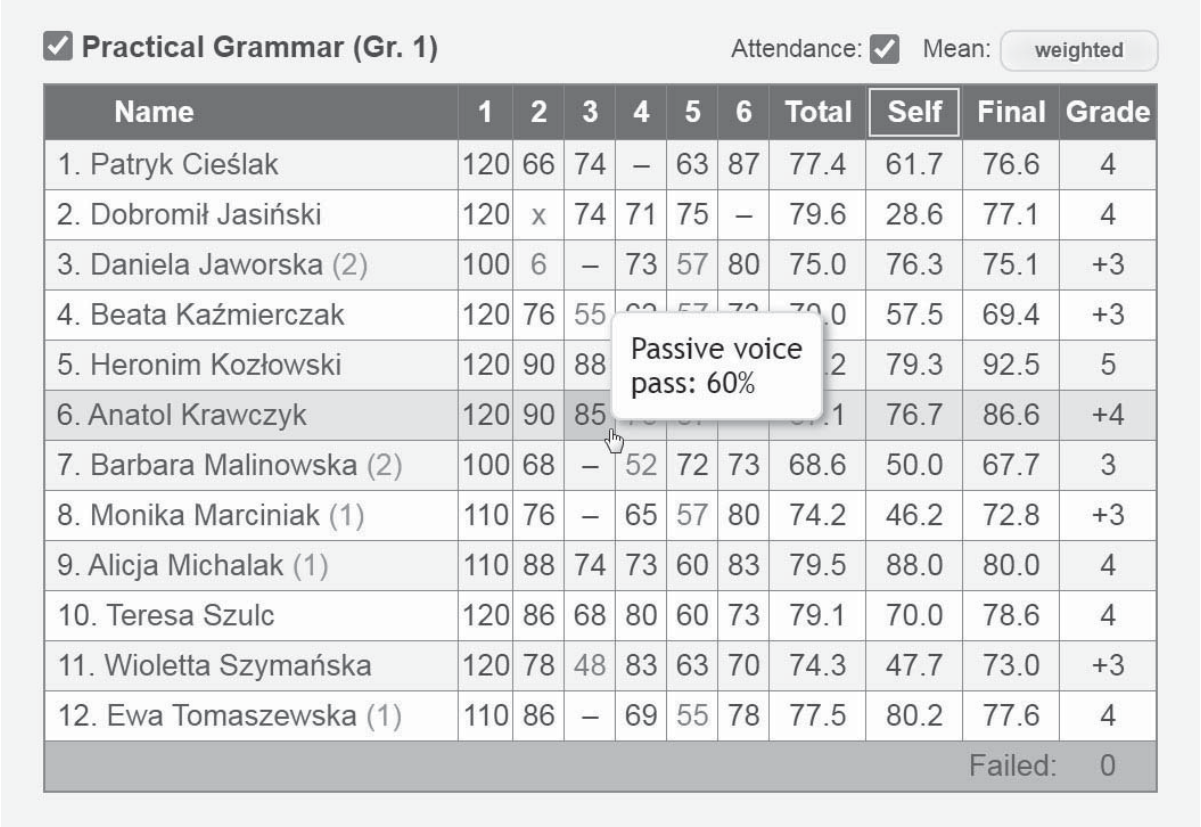
The main part of the report in Figure 2 is a tabulated summary of all the teacher-administered assessments. Attendance can optionally be included in such reports as a component of the overall result, and the precise number of points that students can earn (as well as potential bonus points for having only a few or no absences) can be determined in the settings. For each assessment, the table shows the maximum possible score, the passing score (cut-score), the student's result, and the grade (based on teacher-defined score bands). The grades can be letters (e.g. *A*, *B*, *C*), numbers (e.g. *1*, *2*, *3*), or descriptors (e.g. *excellent*, *very good*, *good*). At the bottom of this table, two means are given, arithmetic and weighted (calculated using the procedures discussed in Section 2.1), and teachers can decide which one is more appropriate for their purposes.

The last part of the report given in Figure 2 is a table that shows the final result and grade, obtained on the basis of the student's self-assessment score and the mean of all the teacher-administered assessments (taken from the main table). This result is calculated by applying weights (specified in the grade settings) given

to these two broad components. However, since self-assessment is not an obligatory constituent of the final result, the teacher may choose not to include it in the reports. In such a case, the last row of the second table in the report is actually left empty, and the results of self-assessment (if any) can be accessed in the teacher's account to be used as feedback on the student's achievement of the learning objectives.

3.3. Reports for teachers

As mentioned above, several score reports are available to teachers. One of them is a tabulated group report which compiles the results of all the assessments administered to an entire class of students. An example of such a report is given in Figure 3.



☑ Practical Grammar (Gr. 1) Attendance: ☑ Mean: weighted

Name	1	2	3	4	5	6	Total	Self	Final	Grade
1. Patryk Cieślak	120	66	74	–	63	87	77.4	61.7	76.6	4
2. Dobromił Jasiński	120	x	74	71	75	–	79.6	28.6	77.1	4
3. Daniela Jaworska (2)	100	6	–	73	57	80	75.0	76.3	75.1	+3
4. Beata Kaźmierczak	120	76	55				0	57.5	69.4	+3
5. Heronim Kozłowski	120	90	88				2	79.3	92.5	5
6. Anatol Krawczyk	120	90	85				1	76.7	86.6	+4
7. Barbara Malinowska (2)	100	68	–	52	72	73	68.6	50.0	67.7	3
8. Monika Marciniak (1)	110	76	–	65	57	80	74.2	46.2	72.8	+3
9. Alicja Michalak (1)	110	88	74	73	60	83	79.5	88.0	80.0	4
10. Teresa Szulc	120	86	68	80	60	73	79.1	70.0	78.6	4
11. Wioletta Szymańska	120	78	48	83	63	70	74.3	47.7	73.0	+3
12. Ewa Tomaszewska (1)	110	86	–	69	55	78	77.5	80.2	77.6	4
									Failed:	0

Passive voice
pass: 60%

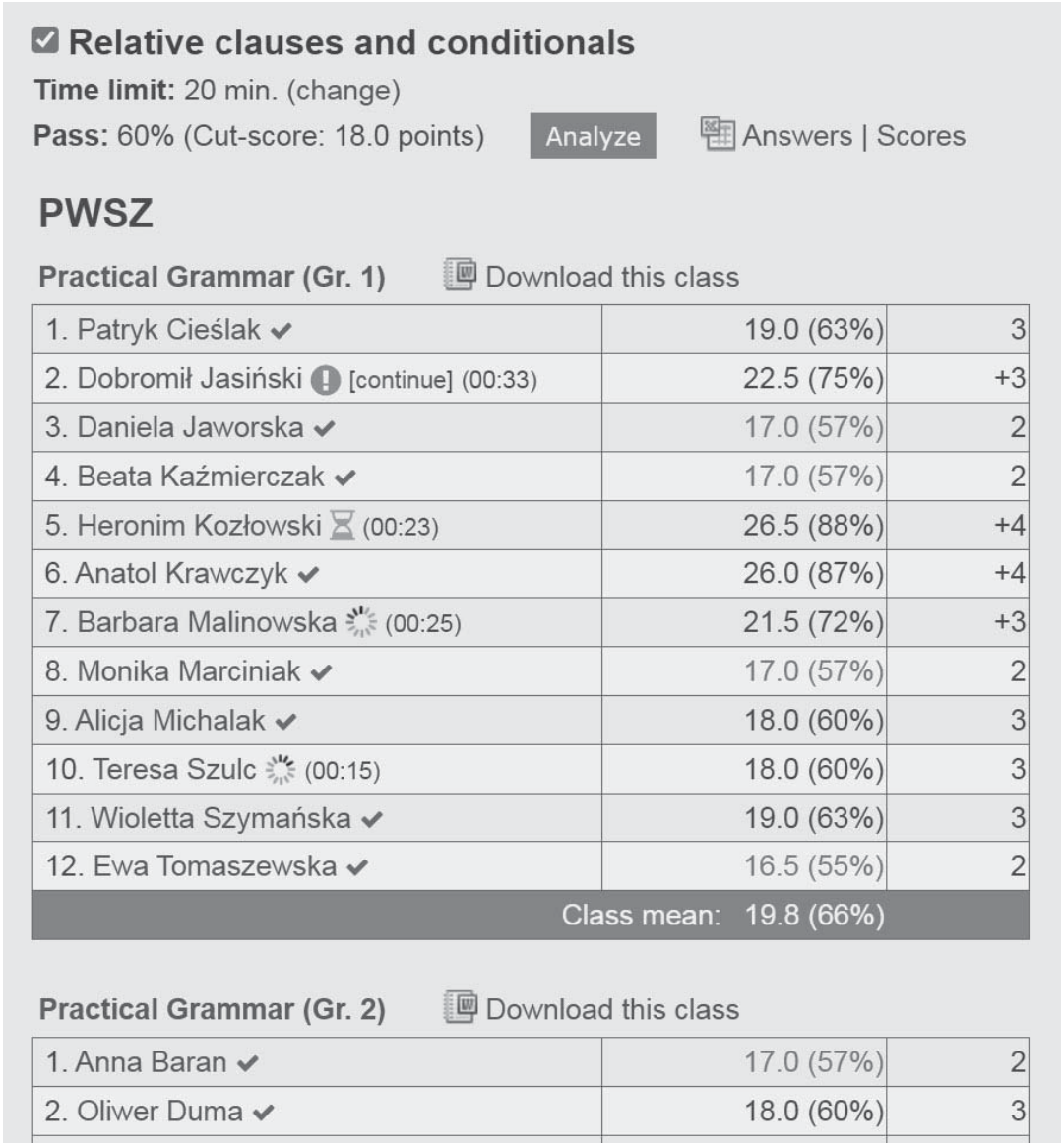
Figure 3. A cumulative score report for a group of students*

* For the purposes of this publication, the original names have been changed using a random name generator.

Each row of the table displayed in a group report is a concise summary of one student's individual cumulative report (which can itself be accessed by clicking the student's name). In this case, however, only percentage scores are given. The assessments are arranged in columns, and each of them is assigned a number in the

header row. As can be seen in Figure 3, scores below the pass mark are displayed in red. The table cells also provide links to the students' individual single-test reports, where the teacher can verify (and make changes to) the automated scoring. Moreover, the report shows the number of unexcused absences (following the names, in brackets). In addition, each column of the table (except for the first one) has an associated tooltip which appears on mouse hover and shows some extra information. Thus, the name of each assessment (and the relevant pass mark) shows up over each numbered column (see an example in Figure 3). The tooltip for the Total column gives information about the number of assessments assigned to each student, the total number of test submissions, and the overall pass mark. The tooltips for the Self and Final columns only show the respective pass marks. The tooltip for the last column shows the cut-off percentage points for the grades.

The second type of score report available to teachers can be found in the test monitoring panel, which comprises a summary of scores, grouped by schools and classes of students. This report, illustrated in Figure 4, includes only those students who have actually started a given test. Thanks to the fact that it is updated in real time, the panel is primarily used to keep an eye on the test-taking process. Further details, such as the time taken to complete the test, the date of submission, as well as the responses, can be accessed by clicking the students' names. Moreover, it is possible to download files containing the students' responses, their scores (including individual item scores and total test scores), as well as entire tests marked automatically (similar to individual single-test reports discussed in the previous section).



Relative clauses and conditionals

Time limit: 20 min. (change)

Pass: 60% (Cut-score: 18.0 points) Analyze Answers | Scores

PWSZ

Practical Grammar (Gr. 1) Download this class

1. Patryk Cieślak ✔	19.0 (63%)	3
2. Dobromił Jasiński [continue] (00:33)	22.5 (75%)	+3
3. Daniela Jaworska ✔	17.0 (57%)	2
4. Beata Kaźmierczak ✔	17.0 (57%)	2
5. Heronim Kozłowski (00:23)	26.5 (88%)	+4
6. Anatol Krawczyk ✔	26.0 (87%)	+4
7. Barbara Malinowska (00:25)	21.5 (72%)	+3
8. Monika Marciniak ✔	17.0 (57%)	2
9. Alicja Michalak ✔	18.0 (60%)	3
10. Teresa Szulc (00:15)	18.0 (60%)	3
11. Wioletta Szymańska ✔	19.0 (63%)	3
12. Ewa Tomaszewska ✔	16.5 (55%)	2
	Class mean: 19.8 (66%)	

Practical Grammar (Gr. 2) Download this class

1. Anna Baran ✔	17.0 (57%)	2
2. Oliwer Duma ✔	18.0 (60%)	3

Figure 4. Part of a test monitoring panel with a summary of scores*

* For the purposes of this publication, the original names have been changed using a random name generator.

Finally, reports summarising the results of various statistical analyses can be accessed from the test monitoring panel (by clicking the 'Analyze' button shown in Figure 4). These reports do not include any students' names: they only provide information about the performance of the individual test items (for item analysis) and the test as a whole (for test evaluation). The details of statistical analysis are rather beyond the scope of this article (see Malec, 2018, for more on that).

4. Conclusions

Test scores are vitally important to all stakeholders, and for this reason score reporting is widely acknowledged as a crucial element of the process of test development. In online education, where technology is used as a means of communication between teachers and learners, appropriate use of digital tools is essential to the provision of accurate and up-to-date information about the students' performance on tests and assessments. Technology should also be seen as a valuable aid because without its assistance many of the activities involved in score reporting would be very time consuming. Moreover, as is often pointed out, humans are prone to making inadvertent errors, due to fatigue, by accident or oversight.

While score reports may be of various kinds, this article concentrated on those which are most likely to be useful in academic contexts, where test scores constitute the basis for making inferences and decisions about the test takers. In the second part of the paper, WebClass has been presented as an online tool which integrates into one all-round system all of the components of test development, including score reporting. It is worth adding that the reports generated on the platform may also include the results of assessments administered elsewhere (cf. the assessment called 'Presentation' in Figure 2), in which case the scores can be entered manually by the teacher. Thanks to this, no other score calculators are needed and, if used appropriately, the platform alone is capable of correctly working out the final semester results (with grades) and reporting them to both teachers and students.

References:

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Bahari, A. (2021). Computer-assisted language proficiency assessment tools and strategies. *Open Learning: The Journal of Open, Distance and e-Learning, 36*(1), 61–87. Doi: 10.1080/02680513.2020.1726738

Brown, J. D. (1996). *Testing in Language Programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, J. D. (2016). Language testing and technology. In F. Farr, L. Murray (eds.). *The Routledge Handbook of Language Learning and Technology* (141–159). London: Routledge.

Brown, J. D., Hudson, T. (2002). *Criterion-referenced Language Testing*. Cambridge: Cambridge University Press.

Carr, N. T. (2011). *Designing and Analyzing Language Tests*. Oxford: Oxford University Press.

Deng, N., Yoo, H. (2009). Resources for reporting test scores: A bibliography for the assessment community. Retrieved from https://ncme.connectedcommunity.org/ncmedev/viewdocument/score-reporting-bibliography

Foreman, S. D. (2018). *The LMS Guidebook: Learning Management Systems Demystified*. Alexandria, VA: ATD Press.

Fuchs, C., Hauck, M., Dooly, M. (eds.). (2021). *Language Education in Digital Spaces: Perspectives on Autonomy and Interaction*. Cham, CH: Springer.

García Laborda, J., Fernández Álvarez, M. (2021). Multilevel language tests: Walking into the land of the unexplored. *Language Learning & Technology, 25*(2), 1–25.

Isbell, D. R., Kremmel, B. (2020). Test Review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing, 37*(4), 600–619. Doi:10.1177/0265532220943483

Kita, Y. (2021). Using Forms quizzes to enhance students' motivation in language learning at a Japanese university. *Indonesian Journal of English Language Teaching and Applied Linguistics, 5*(2), 219–235.

Klimowicz, M. (2020). *Polskie uczelnie w czasie pandemii*. Warszawa: Fundacja Centrum Cyfrowe.

Lane, S., Raymond, M. R., Haladyna, T. M., Downing, S. M. (2016). Test development process. In S. Lane, M. R. Raymond, T. M. Haladyna (eds.), *Handbook of Test Development* (2nd ed., 3–18). New York, NY: Routledge.

Lütge, C., Merse, T. (eds.). (2021). *Digital Teaching and Learning: Perspectives for English Language Education*. Tübingen: Narr Francke Attempto Verlag.

Malec, W. (2015). Fostering autonomous learning with the aid of online technology. In E. Smyrnova-Trybulska (ed.). *IT Tools – Good Practice of Effective Use in Education* (371–381). Katowice: Studio Noa.

Malec, W. (2018). *Developing Web-Based Language Tests*. Lublin: Wydawnictwo KUL.

Malec, W. (2020). Computer-based testing: A necessary evil or a sensible choice? *The Modern Higher Education Review, 5*, 100–113.

Malec, W., Malec, M. (2021). WebClass jako narzędzie do kontroli i oceny postępów w nauce. *Języki Obce w Szkole, 3/2021*, 37–42.

Romaniuk, M. W., Łukasiewicz-Wieleba, J. (2021). Challenges of administering university examinations remotely during the COVID-19 pandemic. *E-mentor, 3*(90), 22–31.

Sendur, A. M., Kościńska, A. (2021). *Kształcenie w sieci – teoria i praktyka*. Kraków: Oficyna Wydawnicza KAAFM.

Soproni, Z. (2020). From item development to analysis: The development of a computerised entrance placement test using a VLE. *Working Papers in Language Pedagogy, 15*, 43–69.

Zenisky, A. L., Hambleton, R. K. (2016). A model and good practices for score reporting. In S. Lane, M. R. Raymond, T. M. Haladyna (eds.). *Handbook of Test Development* (2nd ed. 585–602). New York, NY: Routledge.

(liczba znaków ze spacjami 25 225 + grafika)