

ASSESSMENT AND MEASUREMENT IN HIGHER EDUCATION WITH MULTIPLE CRITERIA, SCALES, AND RATERS: A THEORETICAL APPROACH AND A PRACTICAL EXAMPLE

P. Vossen¹, S. Ajit²

¹*Edumetrics R&D (GERMANY)*

²*University of Northampton (UNITED KINGDOM)*

Abstract

Assessment and measurement (A&M) in higher education can be holistic, relying on the teacher's mental assessment model (covert), or analytic in the sense of computing the sum or average of scores from student's responses on test items or teacher's quality criteria applied to student's course work (overt). Usually, scores are numbers lying on a reduced sort of scale called n-point scale. Although researchers have come up with statistical models to enable A&M with multiple criteria, scales, or raters, the acceptance of those advanced models has remained minimal: they may be too complex for educational practice as they are primarily aimed at educational research projects with enough resources (time, money, people) in appropriate organizational settings.

Since 2003 we have been developing, testing, and applying more appropriate approaches and frameworks for A&M that take the challenging work practices of teachers in higher education into account. They enable A&M with multiple quality criteria on scales of various types, and they allow several assessors to be responsible for distinct quality aspects. The approaches and frameworks were designed with flexibility and adaptability in mind. Often, it is no longer enough to test students for their factual knowledge or general problem-solving ability. Instead, the focus may be on skills that will be assessed through observation of students' behaviour and/or examination of the results of assignments in the form of long-term projects with goals that may be achieved in different but equally valid ways.

Our most recent A&M framework called **Q** provides scoring models for instructional types not yet covered in a satisfying way by traditional A&M, such as group work. We have successfully used **Q** to offer customizable solutions for Group-Peer-Assessment (GPA), deriving individual student scores from a group score and mutual peer ratings, given just two parameters: the spread of student scores around the group score and the impact of peer ratings on student scores. GPA comes with a built-in test called the Split-Join-Invariance (SJI) to check if the final results (marks or grades) are correctly calculated. GPA imposes no artificial restrictions on group size, type or number of quality criteria, or other context-specific aspects of GPA.

Keywords: edumetrics, scores, ratings, marks, grades, bounded scales, scale operations, Group Peer Assessment, peer impact, group score, group spread, peer assessments, student contribution, student score, scoring rule.

1 ASSESSMENT AND MEASUREMENT IN HIGHER EDUCATION

At all levels and in all areas of education, there are three interdependent processes: teaching, learning, and assessing. Teachers teach, learners learn, and assessors assess (outside HE, assessment is often called evaluation or measurement). The role of assessor is traditionally attributed to the teachers, assuming that they are in the best position to assess how well and how much the students have learned. But this may not always be the case. For example, in Group-Peer-Assessment, both teachers and learners play a decisive role in assessment of learning, while teachers may look at other aspects or rubrics than the students. Moreover, when students take part in the overall evaluation of their learning, often a distinction is made between peer assessment (one student assessing another student) and self-assessment (a student assesses his own learning). Of course, the teaching process itself may also be assessed, by the teachers, by the students, or even by a third party who is responsible for higher-level evaluation of all education taking place at an institute. In fact, especially in higher education, assessment and measurement are omnipresent and at least as important as the primary processes of teaching and learning for the mission of educational organizations.

Traditionally, assessment, evaluation or measurement in higher education are either holistic or analytic. It is called holistic if the teacher makes an assessment on a scale that is best considered ordinal. The holistic assessment may be based on several course-specific quality criteria (often in the form of rubrics) that will be applied and aggregated by the teacher into a single overall score. Which quality criteria are applied, and how the aggregation is done need not be made public by the teacher. On the other hand, the analytic tradition is based on calculating the (possibly weighted) arithmetic average of students' scores on a small number of so-called test items, or on a small number of dimensions, i.e., quality criteria applied to students' overall performance on an assignment. Thus, scores are generally conceived of as measures lying on an ordered continuous scales with a small number of discrete anchor points evenly spread over the entire scale.

The holistic type of assessment may be thought of as a black-box type of analytic assessment without test items or quality criteria, and without an explicit rule of aggregation of the component scores. Therefore, we decided to adopt the more general analytic assessment approach – with holistic assessment as an extreme instance of it in which there is just a single composite score and thus no need for a scoring rule. In the case of genuine analytic assessment, researchers have come up with various mathematical models (mostly of a statistical nature) to enable A&M when dealing with multiple criteria and scales. However, those models are too complex for educational practice – witness their non-acceptance as indicated by failing literature on their successful application – but they may be well-suited for large-scale educational research projects with enough resources (money, time, people, expertise), and adequate organizational settings.

Over almost two decades, we have been developing, testing, and applying a more adequate approach to edumetrics, i.e., educational assessment and measurement, that takes the ever more challenging work practices of teachers in higher education into account. Our goal has been to offer educational practitioners, researchers, and scholars a new type of measurement scale fitting most if not all instructional practices, conventions, and traditions. The type of measurement scale shall be as simple as possible and yet powerful enough to solve tough assessment problems such as that of Group-Peer Assessment, i.e., fairly, and consistently grading individual students who collaborated in a group to perform a rather complex project assignment. To enable such challenging assessment tasks, our approach will support customizable quality criteria, on scales of diverse types, potentially applied by multiple assessors, i.e. both teachers and students. Furthermore, our approach will provide explicit (parameterized) scoring rules by which distinct measurements (variously called scores, ratings, marks, or grades) are glued together to yield a single overall score for each student. Finally, we will offer a simple way to check for numerical correctness of the resulting individual scores based on the famous Split-Join-Invariance theorem of GPA.

We decided to keep this paper as non-technical as possible in order to make it accessible and understandable for the practitioner, i.e., teachers of different disciplines who are not primarily interested in the technical details of our assessment approach. Suffice to say, that there is a single mathematical core with just enough components – scales and their operations – to make them work for practice-oriented A&M in a variety of instructional contexts. The mathematical details – including constructions and proofs – can be found in the pioneering paper of Dombi (1) and in our recent publications (1 – 6).

In section 2, we describe the basic structure of our assessment approach and measurement framework dubbed **QI**. We will introduce the notion of bounded scales and show how appropriate operations on scores can be defined. In section 3, we describe how multiple scores can be merged, using the basic operations from section 1, by means of a weighted aggregation function. In section 4, we will apply the concepts introduced in section 2 and 3 to group assessment by the teacher and peer assessment by the students, the first two steps of Group-Peer-Assessment. In section 5, we will explain the role played by the two parameters of any scoring rule which goal it is to merge scores from two distinct scales on a third scale: (1) the impact of one scale on another scale; (2) the spread (or dispersion) of aggregated scores on the target scale. In section 5, the general notion of scoring rule will be introduced: a formula for combining scores on different scales possibly measured by different assessors. Also, we will present the principle of Split-Join-Invariance that can be used to check the correctness of model calculations. In section 6, we will put together the previous components to present a full solution to the Group-Peer-Assessment problem. Finally, in section 7, we will highlight the benefits of our approach to A&M for Higher Education and point out the major implications for teacher training and assessment practice.

2 BASIC STRUCTURE OF THE MEASUREMENT FRAMEWORK ^Q

^Q evolved as a unified framework for working with bounded scales. The concept of bounded scale is quite easy to explain: a bounded scale is a range of numbers called scores with a lower bound **LB** and an upper bound **UB** together with two basic operations called bounded addition $\oplus$ and bounded multiplication $\odot$. Given two scores x and y , we denote their bounded sum by the score $x \oplus y$. Given a score x and a positive number r , we denote their bounded product by $r \odot x$. The number r is not a score, it is called a scalar to remind us of the fact that it is used to “scale” a given score up or down. Hence, the bounded product may also be called a scalar product to distinguish it from the usual product of decimal numbers. The most important feature of bounded addition and multiplication is that all bounded sums and products will again be valid scores between **LB** and **UB**.

Although the bounded sum and product are different from the usual sum $x + y$ and product $r \times x$ of decimal numbers (because of the boundedness of the scale), they share a number of simple properties. For instance, when adding multiple scores, their order and grouping is immaterial for the resulting sum. Also, when multiplying a given score by several positive numbers, the order and grouping of those numbers will have no influence on the resulting product. Furthermore, we will define a neutral score such that when we add it to a given score, that score will not change, and when we multiply it by any given number, we are still left with the neutral score.

The simplest bounded scale might arguably be the scale of logical values **TRUE** and **FALSE** with the logical connectives **OR** or **AND** as bounded addition. It is easy to put the logical connectives into an arithmetic form by mapping **FALSE** to 0 and **TRUE** to 1, and defining bounded addition for Boolean scores x and y , taking on values 0 or 1, as follows:

$$x \oplus_{\text{OR}} y = \frac{x + y}{1 + x \cdot y} \quad x \oplus_{\text{AND}} y = \frac{x \cdot y}{1 + (1 - x) \cdot (1 - y)} \quad (1)$$

Unfortunately, a bounded scale consisting of only two points (Boolean scores) doesn't allow to define a meaningful bounded product, because multiplying a Boolean score with any number doesn't change it at all. However, if we fill up the logical scale so as to allow all numbers between 0 and 1 (i.e., percentages) as bounded scores, then (a) the above formulae for bounded addition still apply, and (b) it is now possible to define bounded, or scalar, multiplication of such scores in the following way:

$$r \odot_{\text{OR}} x = \frac{(1+x)^r \cdot (1-x)^r}{(1+x)^r + (1-x)^r} \quad r \odot_{\text{AND}} x = \frac{2 \cdot x^r}{x^r + (2-x)^r} \quad (2)$$

Although we now have two ways, **OR** and **AND**, of working with bounded scores on the full percentage scale, we are still not quite happy with them, because neither way allows us to effectively represent the notion of compensation between lower and higher scores. What is still missing is the distinction between successes (positive scores) and failures (negative scores) on one and the same scale. The reason is that the **OR**-scale only counts successes while the **AND**-scale only counts failures. Happily, there is a way out by redefining bounded sums and products on the percentage scale as follows:

$$x \oplus y = \frac{x \cdot y}{x \cdot y + (1-x) \cdot (1-y)} \quad r \odot x = \frac{x^r}{x^r + (1-x)^r} \quad (3)$$

This is as far as we will go in this paper with justifying our approach by means of mathematical concepts. Suffice to say here that it turns out to be possible to generalize formulae (3) to all bounded scales with arbitrary endpoints **LB** and **UB** different from 0 and 1, e.g., the signed percentages between **-1** and **+1**, and the well-known Likert scale and other n -point scales (to be converted to the percentage scale).

3 WEIGHTED AGGREGATION OF A SET OF SCORES

The basic structure of the previous section enables us to define the notion of weighted aggregation of a set of scores in a way entirely analogous to the well-known concept of the weighted average of decimal

numbers. Let $\underline{x_1, \dots, x_n}$ be n percentages weighted by the weights $\underline{w_1, \dots, w_n}$ such that $\underline{w_1 + \dots + w_n = 1}$. Using the formulae (3), we define the weighted aggregation $\bar{x}$ of scores $\underline{x_1, \dots, x_n}$ as follows:

$$\bar{x} = (w_1 \odot x_1) \oplus \dots (w_i \odot x_i) \dots \oplus (w_n \odot x_n) \quad (4)$$

In words: calculate the bounded products $\underline{w_i \odot x_i}$ for $i = 1, \dots, n$ and then take the bounded sum of all those products. It can be proved that $\bar{x}$ is again a valid bounded score (here: percentage), and it behaves exactly as we would expect of any decent weighted mean. For instance, if all $\underline{x_i}$ are equal to $\underline{s}$, then $\bar{x} = \underline{s}$, too, and it will never be smaller than the lowest score $\underline{x_i}$ or larger than the highest score $\underline{x_i}$. Most importantly, it is entirely defined in terms of sums and products that make sense for the percentage scale (or any other bounded scale).

4 GROUP ASSESSMENT AND PEER ASSESSMENT

Let us start applying the above concepts to the first two steps of Group-Peer-Assessment by the teacher and the students. Recall that we assume that the teacher will provide an initial group score on the basis of the visible/tangible outcomes of the joint group work. Suppose that there are 5 distinct outcomes (alternatively, they may be thought of as 5 distinct aspects of the joint group product). Let us assume, that the teacher will assess the 5 outcomes or aspects on a percentage scale with 11 anchor points (0%, 10%, ..., 100%). Let's denote the corresponding 5 scores by $\underline{t_1, \dots, t_5}$ with weights $\underline{w_1, \dots, w_5}$ in the aggregated group score $\underline{t}$ (the weights may be equal if there is no reason for differential weighting). Then the following scheme is all that the teacher needs to come up with the initial group score $\underline{t}$:

Table 1. Template for assessment of group work by the teacher

	Weight	Score	Weighted score
outcome 1	$\underline{w_1}$	$\underline{t_1}$	$\underline{w_1 \odot t_1}$
outcome 2	$\underline{w_2}$	$\underline{t_2}$	$\underline{w_2 \odot t_2}$
outcome 3	$\underline{w_3}$	$\underline{t_3}$	$\underline{w_3 \odot t_3}$
outcome 4	$\underline{w_4}$	$\underline{t_4}$	$\underline{w_4 \odot t_4}$
outcome 5	$\underline{w_5}$	$\underline{t_5}$	$\underline{w_5 \odot t_5}$
$\sum_{k=1}^5 w_k = 1$			Group score: $\underline{t} = \oplus_{k=1}^5 w_k \odot x_k$

Now let us do a similar thing for peer assessment. Recall that we assume that the students will co-assess each other on a couple of criteria pertaining to the quality and/or quantity of their contributions to the overall group work, e.g., criteria like cooperation, collaboration, communication, etc. Let us also assume, for the sake of simplicity, that we are not interested in self-assessments (as indicated by neutral score $\frac{1}{2}$). Then each peer $\underline{j}$ will score the students $\underline{i \neq j}$ on $\underline{k}$ criteria using, say, a $\underline{n}$ -point scale, that will be converted to percentages before further processing. For this, they may use a template similar to the one used by the teacher. After all students have done this, we are left with a peer rating matrix with entries denoting the peer ratings $\underline{r_{ij}}$ for student $\underline{i}$ (rows) by peers $\underline{j}$ (columns). Then, the overall student rating for student $\underline{i}$ can be calculated as the (possibly weighted) aggregation over all peers $\underline{j \neq i}$:

Table 2. Template for assessment of students by peers

	Peer 1	Peer 2	Peer 3	Peer 4	Student rating
Student 1	$\frac{1}{2}$	$\underline{r_{12}}$	$\underline{r_{13}}$	$\underline{r_{14}}$	$\underline{r_1} = \oplus_{j=1}^4 w_j \odot r_{1j}$

Student 2	r_{21}	$\frac{1}{2}$	r_{23}	r_{24}	$r_2 = \oplus_{j=1}^4 w_j \odot r_{2j}$
Student 3	r_{31}	r_{32}	$\frac{1}{2}$	r_{34}	$r_3 = \oplus_{j=1}^4 w_j \odot r_{3j}$
Student 4	r_{41}	r_{42}	r_{43}	$\frac{1}{2}$	$r_4 = \oplus_{j=1}^4 w_j \odot r_{4j}$
Group rating:					$\bar{r} = \oplus_{i=1}^4 w_i \odot r_i$

5 GROUP SPREAD AND PEER IMPACT

After the first two steps of group assessment and peer assessment have been done, we need to specify two customization components of the scoring model: the impact of peer assessments on the group score and the spread constraint on the final student scores. These two components of GPA will be modelled as free parameters in the sense that the teacher as principal assessor is entitled to set them in a way that best suits the purpose and constraints of Group-Peer-Assessment in their course.

The first parameter, called group spread (or dispersion) constraint and denoted by z , is used to set an upper and lower limit on the deviation of student scores from the group score $\bar{r}$, thus resulting in a constrained percentage scale. The parameter z can take on values between 0 (no constraint) and 1 (full constraint). If there shall be no constraint, i.e., $z = 0$, the constrained percentage scale will be identical to the full percentage scale, i.e., student scores can take on any value between 0 and 1. The default value of z will be $\frac{1}{2}$, which signifies a moderate level of deviation from the group score $\bar{r}$. If z is set equal to 1, then all student scores will be equal to $\bar{r}$, the group score as set by the teacher.

The second parameter, called peer impact (or peer effect) and denoted by p , limits the extent to which student ratings r_i will have an effect while calculating the final student score using the scoring rule. Often, this parameter is expressed as a percentage, e.g., “peer assessments count for 30%”. However, in our scoring framework, it makes more sense to use a non-negative decimal number with a default value of 1 (standard peer impact). If p equals 0, peer assessments will be discarded, so that all student scores will be equal to the group score. The larger the p is, the stronger will be the effect of peer assessments on final scores, up to the point where student scores will approach either endpoint of the constraint percentage scale – defined by the group spread constraint – depending upon whether the student rating is smaller or larger than the group rating. Due to this phenomenon, the peer impact parameter is sometimes called the polarization parameter.

The two scoring model parameters can be set independently of each other, however, the group spread constraint sets a hard limit to what can be achieved by increasing the peer impact on the final student scores. That means, even for a very high peer impact, student scores will stay within the bounds set by the group spread parameter. Teachers should be careful while setting the parameters off their default values. We recommend starting with the default values until they have accumulated enough experience with real assessment data so they can foresee the effects when changing those parameters.

6 THE SCORING RULE FOR GROUP-PEER-ASSESSMENT

Now we are ready to show the power and usefulness of the signed and unsigned percentage scales with standard neutral scores by using it to solve the problem of Group-Peer-Assessment, i.e., (a) how to define each students' contribution to a group project assignment, and (b) how to use the student contributions to “distribute” a group score over the members of the group in a fair and consistent way.

Question (a) leads to the introduction of the bounded scale $[-1, +1]$ of signed percentages and to an elegant formula for calculating the student contribution in terms of the student ratings, the group rating, and the peer impact parameter (defined above). The individual student contributions will be denoted by signed percentages c_i and defined as follows:

$$(5) \quad c_i = \frac{\left(\frac{r_i}{1-r_i}\right)^p - \left(\frac{\bar{r}}{1-\bar{r}}\right)^p}{\left(\frac{r_i}{1-r_i}\right)^p + \left(\frac{\bar{r}}{1-\bar{r}}\right)^p}$$

Question (b) leads to the definition of a constrained percentage scale $[t/\bar{r}, t^z]$, a subset of the percentage scale around the group score t with the spread constraint z such that the final student scores s_i follow the scoring rule $t/\bar{r} \leq t^{z c_i} \leq t^z$. The constrained percentage scale has all the required properties of a bounded scale in the sense of this paper, i.e., constrained scores can be added and multiplied, while the group score t acts as the neutral score. See Fig. 1.

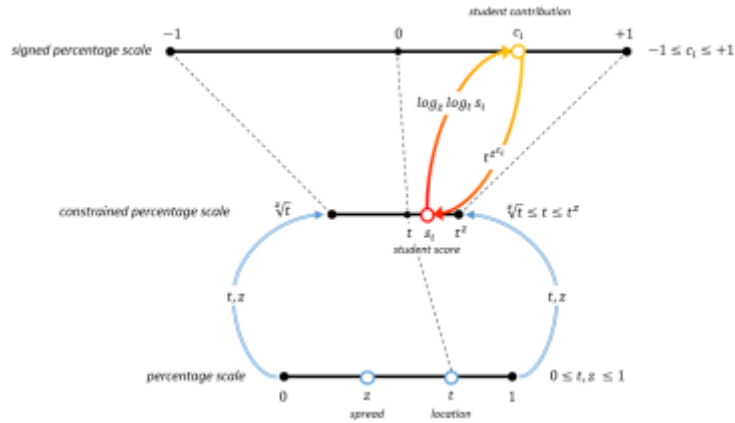


Fig. 1. Overview of Group-Peer-Assessment from a measurement perspective, using a double-exponential scoring function from the signed percentage scale to a constrained percentage scale.

According to the Split-Join-Invariance theorem of Group-Peer Assessment, the weighted aggregation of final student scores s_i equals the group score t (see section 3). Let s_i be an array of n student scores, for $i = 1, \dots, n$, defined by the double-exponential scoring rule $s_i = t^{z c_i}$ with group location score t and group spread constraint z as parameters and the individual student contributions c_i (signed percentages) defined in such a way that the mean contribution $\bar{c} = 0$. Then the weighted mean student score $\bar{s}$, aka computed group score, equals the group location score t for any weighting scheme w_i such that $\sum_{i=1}^n w_i = 1$.

This remarkable theorem provides a practical check of the correctness of GPA calculations: if the bounded mean $\bar{s}$ of weighted student scores s_i is not exactly equal to the original group score t set by the assessor, then we know that something must have gone wrong during the calculations of peer ratings and/or student scores. See Fig. 2, which shows that the group location set by the teacher (left) equals the mean student score (right, highlighted in green):

group location	75%	upper bound	100%	group rating	1.00	group score	87%	group grade	A+
group spread	2.8	maximum	75%		0.55		82%		A+
		mean	58%		0.88		78%		A
		minimum	18%		-0.94		58%		C+
rating impact	3.8	lower bound	0%		-1.00		56%		C

peer matrix: a student in row i being rated by his peers in columns j

	1	2	3	4	5	6	7	8	9	10	student ratings	student contributions	student scores	student grades
1		60%	40%	60%	50%	50%	70%	80%	20%	80%	58%	0.13	77%	A
2	50%		75%	60%	70%	55%	60%	40%	75%	70%	65%	0.37	80%	A+
3	40%	40%		40%	40%	40%	40%	40%	40%	40%	40%	-0.32	70%	B+
4	65%	30%	70%		60%	60%	40%	90%	60%	30%	58%	0.21	75%	A
5	40%	38%	60%	20%		60%	60%	60%	70%	20%	44%	-0.18	72%	A-
6	80%	40%	40%	60%	80%		80%	80%	60%	60%	66%	0.38	80%	A+
7	60%	70%	65%	70%	80%	50%		80%	75%	70%	76%	0.55	82%	A+
8	60%	35%	40%	30%	65%	30%	20%		40%	30%	36%	-0.45	68%	B+
9	50%	50%	50%	50%	50%	50%	50%	50%		50%	50%	0.00	75%	A-
10	60%	10%	10%	10%	10%	10%	10%	10%	50%		18%	-0.94	58%	C+

Fig. 2. Template for setting assessment parameters, collecting the peer assessments, and calculating student ratings, student contributions, as well as student scores and grades.

7 BENEFITS AND IMPLICATIONS

Our assessment framework offers several benefits for teachers and students alike:

- (1) It offers a unified vocabulary for talking about assessor roles and assessment components.
- (2) It can be adapted to different curricula, courses, and classes.
- (3) It provides a simple meta-structure to specify the components of assessment and measurement in a given context.
- (4) The framework components can be re-used or further specified or modified.
- (5) Quality criteria in framework Q are not built-in (fixed). What counts is that they are strongly tied to the course-specific assignment or test and may have distinct measurement scales.
- (6) Our framework allows many types and forms of scales, such as n-point scales or (signed) percentage scales.
- (7) It refrains from specifying a fixed grading scale because the latter differs substantively between countries.

There are three main implications for teacher training and assessment practice:

- (1) Preservice teachers need to get up-to-date training in A&M methods and tools, incl. frameworks like Q. Once in service, it is too late to pick up the necessary vocabulary and skills.
- (2) On the job, junior teachers will get sufficient support from management and staff to acquire new assessment and measurement tools and to use or evaluate them in their class.
- (3) Students need instruction and training in modern assessment and measurement concepts and techniques to understand and appreciate alternative ways of A&M in higher education.

Earlier versions of our GPA framework have been used as an Excel VB application. We are working on the integration of GPA into existing peer assessment tools, and on the development of an online software app using Google Tables in combination with Google Apps Script.

REFERENCES

- [1] J. Dombi, "Basic concepts for a theory of evaluation: The aggregative operator," in Eur. J. Oper. Res., vol. 10, no. 3, pp. 282–293, 1982.
- [2] P. H. Vossen, S. Ajit and A. Dean, "Quasi-arithmetic scoring and group project marking," in Proceedings of the ICERI2019 conference, 2019.

- [3] P. H. Vossen and S. Ajit, "Fuzzy scoring theory applied to team-peer assessment: Additive vs. multiplicative scoring models on the signed or unsigned unit interval," in V. E. Balas et al. eds., *Soft Computing Applications*, Springer International Publishing, pp. 84–111, 2020.
- [4] P. H. Vossen and S. Ajit, "Towards a Paradigm Change in Group and Peer Assessment in Software Engineering Education", in *IEEE 32nd Conference on Software Engineering Education and Training*, pp. 1-10, 2020.
- [5] P. H. Vossen and S. Ajit, "The Challenge of Peer Assessment in Higher Education", in *ICERI 2022 Proceedings, IATED*, pp. 8603-8612, 2022.
- [6] P. H. Vossen and S. Ajit, "Bounded scales: a new type of scale for educational practice and research", In *EDULEARN23 Proceedings*, pp. 6912-6922, 2023.
- [7] P. H. Vossen and S. Ajit, Group-peer-assessment from a quantitative point of view. In *EDULEARN23 Proceedings*, pp. 6979-6989, 2023.