Fuzzy Scoring Theory applied to Team-Peer Assessment: *additive vs. multiplicative scoring models on the signed or unsigned unit interval*

Paul Hubert Vossen[1,(✉*] **and Suraj Ajit**[2]

[1] Independent Researcher, Kiefernweg 1A, D-97996 Niederstetten, F.R. of Germany, Phone: +49 (7932) 3678698

[2] Northampton University, UK Department of Computing, Faculty of Arts, Science and Technology, Newton Building, St. George's Avenue, NN2 6JB, The University of Northampton, Phone: +44 (0)1604 893257

✉ p.h.vossen@googlemail.com

Abstract. Teamwork in educational settings for learning and assessment has a long tradition. The reasons, goals and methods for introducing teamwork in courses may vary substantially. However, in the end, teamwork must be assessed at the group level as well as on the student level. The lecturer must be able to give students credit points or formal grades for their joint output (product) as well as for their cooperation in the team (process). Schemes for such multicriteria quantitative assessments appear difficult to define in a plausible way. Over the last five decades, plenty proposals for assessing teamwork processes and products on team and student level have been given using diverse scoring schemes. There is a broad field of empirical research and practical advice about how team-based educational assessment might be set up, implemented, improved, and accepted by staff and students. However, the underlying methodological problems with respect to the merging of several independent measurements has been severely underestimated. Here, we offer an entirely new paradigm and taxonomy of teamwork-based assessment following a rigorous fuzzy-algebraic approach based on two core notions: quasi-arithmetic means, and split-join-invariance. We will show how our novel approach solves the problem of team-peer-assessment by means of appropriate software tools.

Keywords: Performance Assessment Scoring Systems, Team-Peer-Assessment, Collaborative Learning, Learning Groups, Scoring Algebra, Additive Scoring, Multiplicative Scoring, Quasi-Arithmetic Means, Split-Join-Invariance, Scoring Function, Scoring Equation, Peer Rating, Student Scoring, Zooming Factor

1 Introduction

Team-Peer-Assessment (TPA for short) has a long tradition (cf. Dochy et al. 1999; Falchikov 1986; Falchikov 1993; Falchikov & Goldfinch 2000; Gibbs 2009; Strijbos & Sluijsmans 2010; Topping 1998), Van Rensburg 2012). One of the most cited approaches has been described in a paper by Sharp (2006) (see section 2). Recently, research practice has moved from assessment model development to empirical studies focussing on didactic, cognitive and social aspects of the assessment of teamwork (cf. Dijkstra et al 2016).
However, it would be false to conclude that the problem of deriving individual student marks, scores or grades from corresponding measures on team level has already been solved adequately and completely. On the contrary, most TPA practices still rely uncritically on scoring proposals following statistical and similar approaches which don't consider that educational assessment practices almost invariably work with two-sided bounded scales quite different from the real numbers used in statistics (cf. Sharp 2006). Other proposals follow standard psychometric approaches like Item Response Theory (e.g., Ueno 2010; Uto & Ueno 2016) which are well-suited for educational research but not for everyday practice in the classroom.

Therefore, we started in 2007 a research project aimed at developing a measurement approach for educational assessment which is geared to the needs of practising lecturers without sacrificing formal requirements such as correctness, completeness, flexibility, and simplicity. Our approach is based on concepts, methods, techniques and tools borrowed from diverse fields, notably fuzzy mathematics (e.g. Fodor 2009; Dombi 1982), measurement theory (e.g. Batchelder et al. 2016), functional equations (e.g. Aczél 1966; Ng 2016), and the theory of quasi-arithmetic means, a powerful generalisation of the common arithmetic mean (Kolmogorov 1930; Jäger 2005).

Previous results have been presented at numerous conferences and published in the conference proceedings, e.g. (Vossen & Kennedy 2017a/b, Vossen 2018). Here we will present our most recent results regarding the shift to such a rigorous educational measurement paradigm for TPA (see also Bukowski et al. 2017). In a nutshell, the fundamental TPA problem can be formulated as follows:

> *How to assign correct and fair individual scores to the members of a learning team, if the only judgmental data we have are an overall team output score assigned by the lecturer and the mutual peerwise ratings of team dynamics by the students (excl. self-assessments)?*

Although the same approach may also be used outside the TPA paradigm, e.g. when the lecturer himself provides all judgmental data or when assessment is fully integrated in an e-learning context, we will not further elaborate on this here.

Likewise, empirical studies about the acceptance, dissemination and actual use of the new TPA ideas presented here are not the topic of the current paper, but we encourage educational practitioners to reflect seriously on their own practice of TPA and compare it with the approaches and models we propose here.

We will show that a full theory and method of TPA can be developed bottom-up from a few core concepts, constructs and principles, notably the concept of *Quasi-Arithmetic Mean* (see glossary) and the principle of *Split-Join-Invariance*. In a nutshell, the latter principle states:

> *If the overall team score as judged by the lecturer based on a course-dependent list of assessment criteria is split into separate, possibly different student scores using a predefined scoring function, then the corresponding quasi-arithmetic mean of the individual student scores shall be equal to the initial lecture score.*

Here is a graphical illustration of this principle:

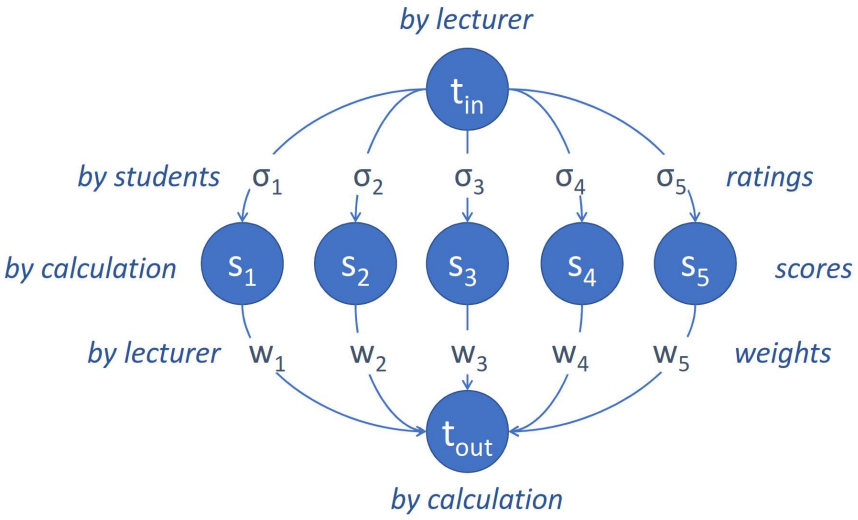

Figure 1: Illustration of the Split-Join-Invariance principle (see text and glossary for explication)

It turns out that there are four distinct scoring functions or rules making up a two by two taxonomy of assessment schemes or models: signed additive ($A^{\pm}$), unsigned additive (A^{+}), signed multiplicative ($M^{\pm}$), and unsigned multiplicative (M^{+}). The following figure gives a high-level description of these four scoring rules. Full details and derivations will be given in section 4. A fully worked out representative assessment for a team of five students using one of the four scoring models will appear in section 3. For a quick look-up and short explanations of the main concepts of TPA, see also the alphabetic glossary at the end of this paper.

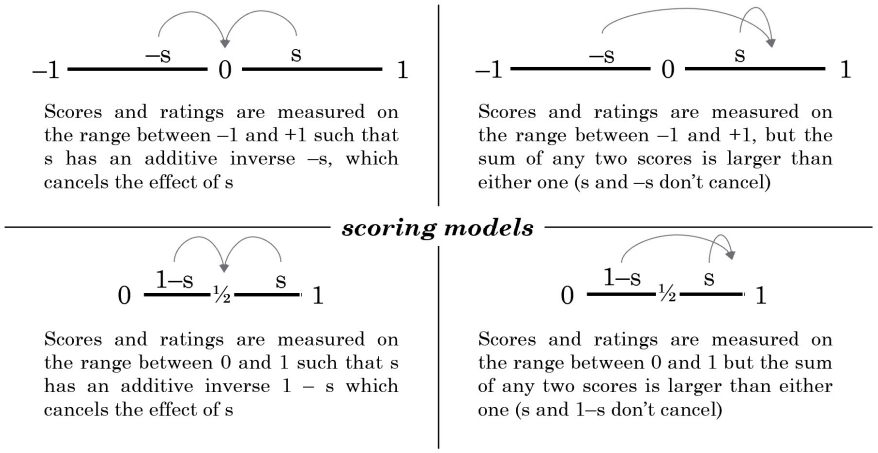

Figure 2: The two by two taxonomy of TPA scoring models

However, before we delve into this novel theory of Team-Peer-Assessment let us give a reconstruction of two classical approaches in which scores (used by the lecturer to assess the team's product) and ratings (used by the students to assess the team's process) are incorrectly treated as any real numbers endowed with ordinary arithmetic addition and multiplication instead of fuzzy-mathematics based quasi-arithmetic addition and multiplication of scores or ratings on bounded ranges.

2 Two reconstructions ~ shorter

The linear statistical approach of Sharp

The linear statistical approach of Sharp (2006, p. 335) which uses well-known statistical techniques of analysis of variance (ANOVA) is based on a simple assumption about the relationship between marks (percentages) and peer ratings (unspecified) (p. 341):

$$(1) \qquad \left(M_j - \bar{M}\right) = \varphi \times \left(S_j - \bar{S}\right)$$

Here, M_j is the final mark of a student, $\bar{M}$ is the so-called location factor set equal to the tutor mark (as a global average for the entire team, see below), S_j is the contribution of student j to the joint group work, and $\bar{S}$ is the (arithmetic) mean of all peer ratings. The constant of proportionality φ (Sharp's terminology) is an arbitrary

rescaling factor "to be determined in the light of empirical evidence" (p. 333). What kind of empirical evidence is meant here and how a choice can be justified in the light of it is not fully clear from the paper. Note however that in the case of (too) large deviations from $\overline{M}$ and/or $\overline{S}$, it may readily happen that $M_j > 100$, the given upper limit of marks (p. 335: "The scales of linear statistical models… are not constrained within particular intervals").

A substantial part of the appendix to Sharp's paper (p. 341-343) is devoted to the calculation of a statistical measure of detectability (A) which shall be used to decide whether to moderate tutor's overall score ($\overline{M}$) at all or not. If A is such that the tutor may conclude that the differences in rating between the students are not significant (in statistical terms) then the recommendation is not to moderate ("… 'switch the system off' for that group.", p. 338). This aspect of the proposal is highly technical, as Sharp acknowledges himself (p. 337): "The method described here is rigorous but not easy to explain to students unfamiliar with analysis of variance.". We hasten to add, that (this part of) the method may also appear prohibitive for tutors or lecturers from disciplines in which empirical research based on (advanced) statistical techniques is not in their usual bag of skills. At the end of his paper, Sharp warns, that because of the statistical approach his method only works reliably for teams of size four or more: small teams of 2 or 3 students are ignored at all.

It will be clear from our summary that we have strong doubts about the viability of Sharp's proposal. On the one hand, it uses a *linear* scoring model and formula which are not appropriate for the type of measurement scales used for scores and ratings and unnecessarily relies too much on the creativity and intuition of the assessor. On the other hand, it introduces research techniques that may well be adequate for empirical researchers, but that may overwhelm the intended population of practising lecturers and students in practice-oriented disciplines (think e.g. of language courses, design curricula, and practical medical, legal or social studies).

2.2 The quadratic numerical approach of Nepal

The paper (Nepal 2012) is relevant because of its unusual approach to the issue of how much the final student score may deviate from the score given by the lecturer. Nepal proposes a peer rating model which substantially differs from the prototypical approaches found elsewhere because it breaks with the dominant linear statistical tradition which culminated in the proposal in (Sharp 2006). Unlike Sharp's, his basic equation is non-linear in the peer assessment influence on lecturer's score. It is a quadratic polynomial with two parameters. We will start with a reconstruction, pointing out its good ideas, and then conclude why we believe that this approach nonetheless does not enable a real breakthrough a.o. because of its very specificity.

Nepal uses an *individual weighting factor* IWF which is also implicit in the default solution for φ proposed by Sharp (2006: p. 342) except for how the average contribution is calculated:

$$\text{IWF} = \frac{\text{Individual contribution (\%)}}{\text{Average contribution (\%)}} = \frac{\text{Individual contribution (\%)}}{(100/n)} \tag{2}$$

Nepal considers the determination of the individual weighting factor IWF as non-problematic. The determination rests upon an estimate of the individual contribution to the project using co-assessments (p. 555), and a fixed average contribution of 100/n, where n is the team size (p. 566).

However, Nepal suggests an entirely different way of adjusting the team mark TM (on a scale from 0 to 100) by IWF (a non-negative real number). After shortly reviewing four well-known formulae for calculating an individual mark IM on the base of TM and IWF, Nepal introduces his so-called *parabolic formula* (p. 557):

$$IM = TM \times \begin{cases} IWF & [IWF \leq 1] \\ IWF - \frac{(IWF-1)^2}{2\alpha\left(1-\frac{TM}{100}\right)} & \left[1 < IWF < 1 + \alpha\left(1 - \frac{TM}{100}\right)\right] \\ 1 + \frac{\alpha}{2}\left(1 - \frac{TM}{100}\right) & \left[IWF \geq 1 + \alpha\left(1 - \frac{TM}{100}\right)\right] \end{cases} \tag{3}$$

Here, IM is the mark awarded to an individual team member; IWF is the individual weighting factor; TM is the team mark, and α is a scaling parameter. This parameter $\alpha \geq 0$ is introduced to enable adjustment of the impact of IWF to local "mark-grade translations" (his phrase, p. 557), but it also plays a role in enforcing the resulting individual mark to stay within the allowed range from 0 to 100 (p. 558).

As can be seen from the definition of IM (individual mark) in formula (3), Nepal distinguishes three regions of different behaviour of this function: up to 1, between 1 and 1 + α(1–t), and above 1 + α(1–t). For w up to 1, IM is independent of t and α. From 1 + α(1–t) upward, IM will be equal to 1 + ½α(1–t), the maximum value of the parabolic function defined in (3) for the region between 1 and 1 + α(1–t). The regions and behaviour of the function are chosen such that there are no gaps (*discontinuities*) at the transition points 1 and 1 + α(1–t).

Here is an example of the calculation of student score IM for TM = 80, w = 2, and α =5. As w = 2 = 1 + 5 × (1 – 0.8), it follows that the third case of formula (3) applies: 1 + ½ × 5 × (1 – 0.8) = 1½. Since TM = 80, this yields an individual mark of $80 \times 1½ = 120$. Obviously, this student has done quite well. Moreover, due to the high values of both w and α, the calculated individual mark of 120 is beyond the maximum allowable mark of 100. The "creative solution" is to set the final individual mark to *min*(100,120) = 100.

Nepal's proposal is in certain respects an improvement upon previous peer rating or co-assessment proposals. First, Nepal defined explicitly a standard average contribution to a project team which only depends on the team size (n), if the total workload will be evenly distributed and taken up by all members of a cooperative and well-coordinated team. In a sense, it is like criterion-based testing instead of norm-based testing, where the criterion is "equal contribution by all team members in percentages of the required workload" and the norm would have been "average contribution of all team members as indicated by the peer assessment reports".

Unclear is what should be changed in the calculation, if the team does not deliver according to the total required workload (team underachievement) or if the team delivers more than required (team overachievement).

Second, Nepal courageously departs from the usual linear-statistical thinking style of the mark or score adjustment, which uses something of the form $a \cdot t + b$, where t stands for a team score, and a and b are some formally or empirically determined modifiers or parameters, e.g. peer ratings as discussed here. Instead, he suggests, that this way of adjusting the team mark is only OK for individual weightings up to one. However, for individual weightings above this standard, or default value of 1, the effect of the weighting upon the team score should be deflated, to avoid overambitious team members to "eat up" all the work instead of co-operating with their peers (the reverse of free-riding, so to speak.

The scaling parameter is used to adjust the individual weightings to be in line with local marking policies. However, Nepal points out, that only a scale factor of fewer than two guarantees that the adjusted mark will fall within the allowable range from 0 to 100. In fact, the condition is that the product of the scaling factor and the team score is less than two, but that does not change the line of argument. Thus, the problem of out-of-range scores is not eliminated at all. The underlying cause is that polynomial functions are not appropriate for simultaneously enforcing an upper and lower bound on the values of a function unless one arbitrarily restricts the range of the function variable.

3 A gentle end-user's introduction to the TPA approach

Let's start with a realistic case study of peer assessment for a team of five students using a standard spreadsheet tool (here, we have used Excel, but any other modern spreadsheet software would do as well, e.g. google sheets). For the moment, we don't care about which of the four standard models will be used: for the end-user (lecturer or student), all would appear exactly the same, except of course for the scores or ratings that are allowed, as that depends upon whether the lecturer uses a signed [-1,+1] or unsigned [0,1] scale.

Let's assume the lecturer starts with inputting his judgments: the team product score $t_{lecturer} = 0{,}70$ (+70%), the default zooming factor $z_{lecturer} = 1{,}0$ (decreasing or increasing the impact of student ratings on the team score) as well as the individual student weights $w_1, \ldots w_5$:

Team level	**Signed Additive Scoring**					*by lecturer*
$z_{lecturer}$	w_1	w_2	w_3	w_4	w_5	$t_{lecturer}$
1,0	0,20	0,10	0,15	0,20	0,35	0,70

Figure 3: Data input of the lecturer: zooming z, weights w and team score t.

As can be seen, there were some issues with students 2, 3 and 5, the others take their usual share of 1/5 = 20%. Students 2 and 3 may have been absent or sick for a short period of time, or there may be other reasons why they contributed less to the project than formally required or expected. On the other hand, student 5 apparently took over the work which was not done by students 2 and 3. The weights as such don't tell us anything about the quality of the work, it's just a quantitative indication of the participation or involvement of a student in the teamwork due to external factors.

The zoom factor or zoom parameter z has been set by this lecturer on its default value of 1, which means that the lecturer has initially no reason to manually shrink or stretch the range of final student scores. That is, he or she will be confident that the students will do their best to come up with differential peer ratings so that systematic differences in cooperative behaviour of peers will be captured and have corresponding impacts on their final scores. The zooming factor may later be adjusted if this initial hypothesis about correct and fair peer rating turns out to be somehow wrong or not plausible.

Now it's time for the students to deliver their judgments of each other's cooperative behaviour during teamwork, which is a pragmatic way to capture team dynamics. Each team member will rate each other team member on a list of process quality criteria using the same scale which will be used for all other judgments (this simplifies the scheme but is not really necessary). The ratings for one and the same peer (5 – 1 = 4, i.e. excluding self-assessment) will be averaged using the quasi-mean belonging to the scale. This rating will be recorded in the so-called peer assessment matrix (see Figure 4):

Peer level **Signed Additive Scoring** *by team*

p_{ij}	p_1	p_2	p_3	p_4	p_5	
p_1	–	0,70	0,65	0,90	0,70	peer 1 assesses peers 2, 3, 4, 5
p_2	0,50	–	0,70	0,90	0,60	peer 2 assesses peers 1, 3, 4, 5
p_3	0,55	0,65	–	0,75	0,70	peer 3 assesses peers 1, 2, 4, 5
p_4	0,60	0,70	0,60	–	0,70	peer 4 assesses peers 1, 2, 3, 5
p_5	0,45	0,60	0,80	0,85	–	peer 5 assesses peers 1, 2, 3, 4

Figure 4: Mutual peer ratings of all members of the learning team

Given that we are working here with a signed additive scoring scheme the students appear to be very satisfied with each other's cooperation in the team. Still, there are (relatively small) differences, but no systematic outliers. How strong will this impact the final student scores which can now be calculated? Here are the numbers:

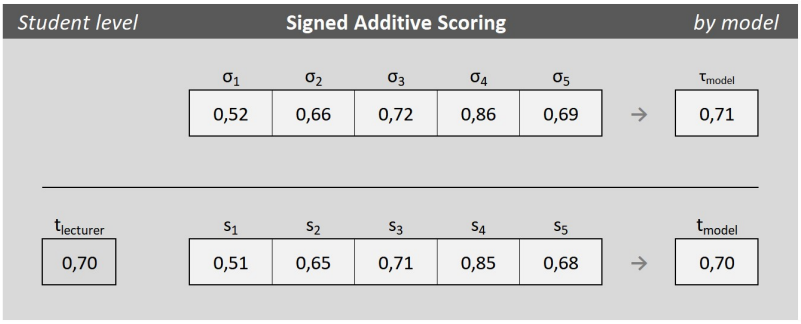

Figure 5: Calculated student ratings (σ_i), mean rating (τ), student scores (s_i) and mean score (t)

We happily observe that the individual peer ratings are indeed spread out clearly, from 0,52 (σ_1) to 0,86 (σ_4), so there seems to be no reason to adjust the zooming factor. Their impact on the team score of 0,70 set by the lecturer is correspondingly clear. We see this immediately in the following diagram:

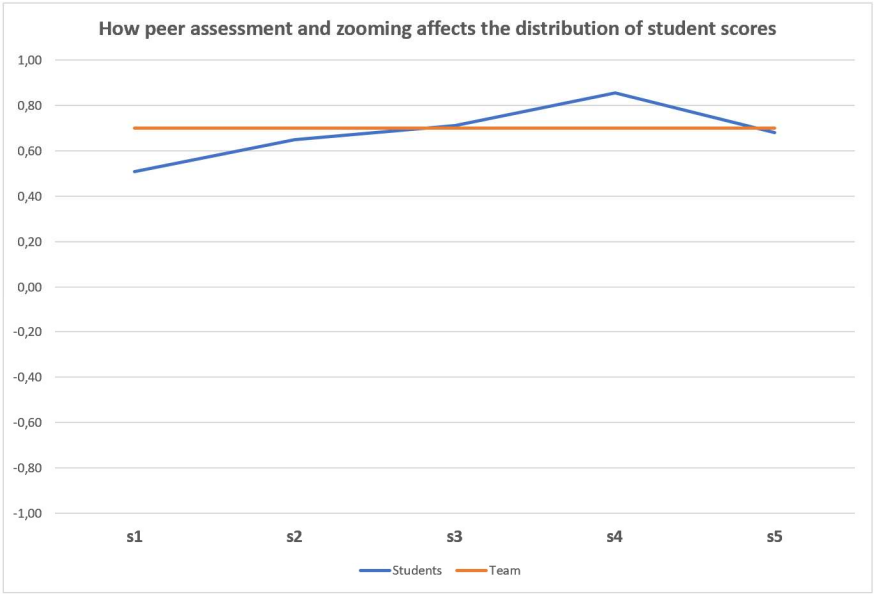

Figure 6: Peer adjusted student scores around a team score of 0,70 set by the lecturer for a default zooming (impact) factor of one

Nevertheless, for didactic purposes, let's increase the zooming factor to two so that we can explore its effect, or impact, on the dispersion (i.e., differentiation) of the final student scores:

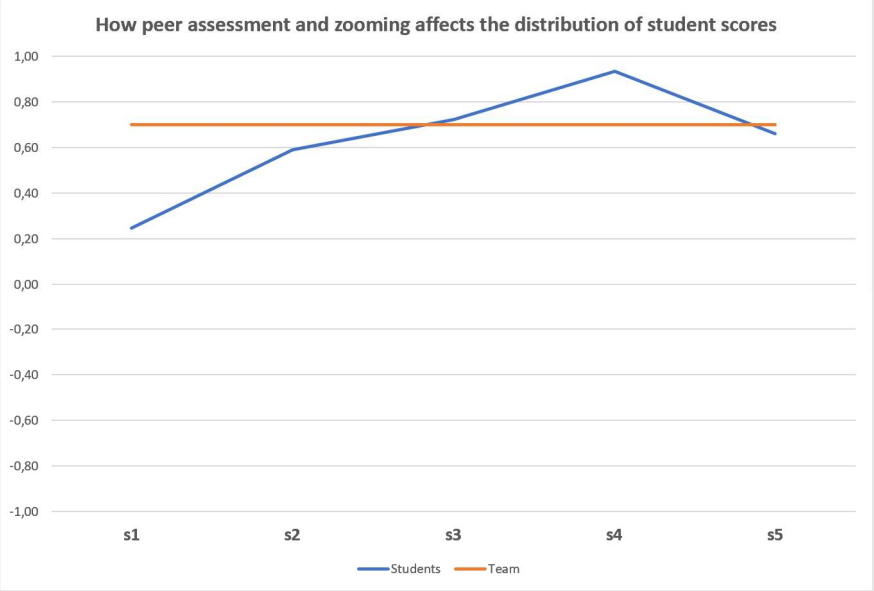

Figure 7: Peer adjusted student scores around a team score of 0,70 set by the lecturer for an increased zooming factor of two

Now the scores range from 0,25 to 0,93, which will very probably have a noticeable impact on the grades that will be calculated from these scores (grades will be no issue in this paper, as the types of grading systems worldwide are too diverse to be considered in our system). Note, however, that all scores are still above the pass-fail threshold of 0 on the signed additive scale [-1,+1] that we are using here.

Finally, it is very important to point out that the principle of Split-Join-Invariance (SJI) has been satisfied (cf. Figure 1, glossary). This can be clearly seen in Figure 4, where $t_{lecturer}$ and t_{model} are exactly equal, as required by SJI. If the team average calculated by the model (t_{model}) would not be equal to the initial global team score given by the lecturer ($t_{lecturer}$), then students would have reason to doubt the correctness of the assessment procedure: either the model didn't take the lecturer's global score into due account or the lecturer didn't take all the available evidence about teamwork into account. This is the very reason why we formulated and introduced the SJI.

4. Basics of the scoring models

The TPA tool introduced in the preceding section rests on the following conceptual model:

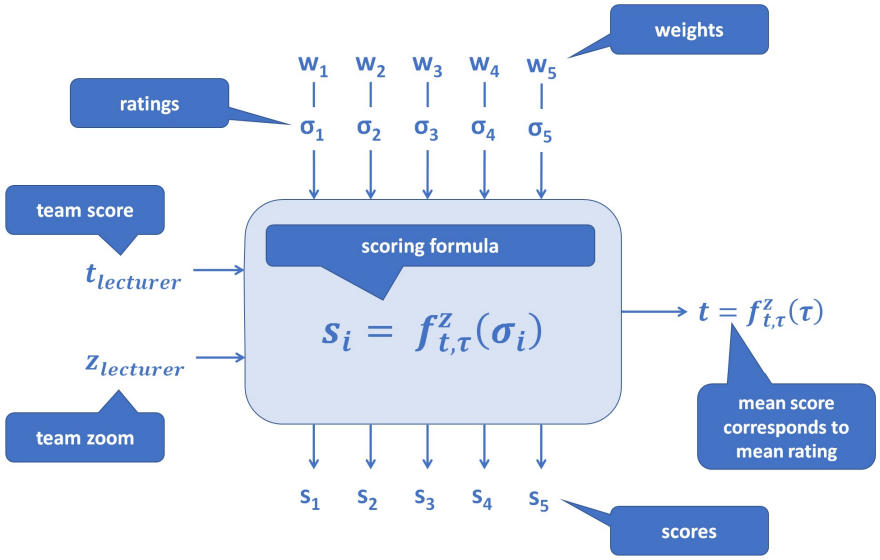

Figure 8: Input-output diagram of our TPA scoring models

The core concept is the scoring formula or function $f^z_{\tau,t}(\sigma_i)$ which transforms peer ratings σ_i into student scores s_i. Evidently, this function needs three parameters or modifiers to work:

- The team product score $t_{lecturer}$. This team score is initially set by the lecturer based on the team's products (outputs, deliverables) of any type relevant for the original task assignment, e.g. a paper, a design prototype, software or hardware, the solution of a mathematical problem, translation, presentation, demonstration, etc. Usually, this product score will be calculated as the mean of several quality factors or criteria using the adopted scoring scale (signed or unsigned, additive or multiplicative). The default values are scale-dependent, 0 or ½.

- An initial team zoom factor $z_{lecturer}$. This zooming parameter determines the spread of final student scores around the team product score. It is a non-negative real number, with a default setting of 1, which gives the simplest scoring function. To get more differentiation, as perhaps required by faculty admin, the scores may be spread (stretched) by setting $z > 1$; setting $z<1$ reduces the spread (shrinking), e.g. if there seems to be adverse coalition forming in the team. With $z = 0$, peer assessment will effectively be disabled, e.g. in the extreme case that apparently the students didn't understand or follow the rules of the (assessment) game or if their team dynamics was ostensibly completely out of order.

- The team rating τ, i.e. the mean of the individual student ratings. This quasi-mean can be calculated straightforwardly once the student ratings have been calculated from the peer matrix. It needs the student weights, which are non-negative real numbers summing to one.

Once the final student scores have been calculated by means of the scoring function, it is an easy matter to calculate the mean score for this team, again using the student weights set by the lecturer (as before for the team rating). Due to the SJI principle, this calculated team score shall be equal to the initial team score given by the lecturer. If not, there is something wrong either with the entered data (invalid data input, e.g. out of range), with the parameter settings (e.g., invalid parameter values) or with the spreadsheet (implementation errors, e.g. lost or false functions).

The only component which is still undefined at this point in the discussion is the scoring function $f_{\tau,t}^{z}(\sigma_i)$. How shall we proceed? In hindsight, it is quite simple. To find adequate scoring functions, we have constructed four scoring algebras. Here are the basics:

(1) analogously to the real line, we have endowed the *additive scoring scales* with suitable operations of addition, subtraction, and scalar multiplication with a positive number (division is missing, we don't need it);

(2) analogously to the positive real line, we have endowed the *multiplicative scoring scales* with suitable operations of addition, scalar multiplication with a positive number and scalar division yielding a positive number (subtraction is missing, it is not available).

The important thing about these operations is the following. When one applies addition or subtraction to two scores or ratings one again gets a valid score or rating within the original range. Scalar multiplication and scalar division are somewhat different. With scalar multiplication, a non-negative number is applied to a score or rating, yielding again a valid score or rating. With scalar division, one score or rating is divided by another score or rating to yield a non-negative number, i.e., a scalar, which - when applied to the latter score or rating - would again yield the former score or rating. [*It would be possible to define genuine multiplication and division though at the expense of getting more complicated formulae. We hope to find another way to handle scalar division, but for the moment it works fine*].

Obviously, the operations we have just introduced will *not* be the usual arithmetic operations of addition, subtraction, multiplication and division on the real line. To remind us of this fact and to be able to distinguish between ordinary arithmetic operations and the new quasi-arithmetic operations we have chosen proper symbols for them which remind us of their corresponding arithmetic ones:

- $x \oplus y$ for quasi-addition, defined in all models
- $x \ominus y$ for quasi-subtraction, defined in both additive models
- $r \odot x$ for quasi-multiplication, defined in all models
- $x \oslash y$ for quasi-division, defined in both multiplicative models

With these operations, we can define quasi-arithmetic means (quasi-means for short) in all models, where t is the team score and τ is the team rating:

(4) $t \stackrel{\text{def}}{\equiv} \bigoplus_{i=1}^{n} w_i \odot s_i$ t is the weighted quasi-sum of the s_i

(5) $\tau \stackrel{\text{def}}{\equiv} \bigoplus_{i=1}^{n} w_i \odot \sigma_i$ τ is the weighted quasi-sum of the σ_i

Finally, we are ready to give a concise definition of the additive and multiplicative scoring functions in terms of scores, mean score, ratings, mean rating, zoom factor and the quasi-operations we have introduced before:

(6) $$s_i \stackrel{\text{def}}{\equiv} f_{\tau,t}^{z}(\sigma_i) \stackrel{\text{def}}{\equiv} \begin{cases} t \oplus [z \odot (\sigma_i \ominus \tau)] & \text{for additive scoring} \\ (\sigma_i \oslash \tau)^z \odot t & \text{for multiplicative scoring} \end{cases}$$

Expressed in the usual arithmetic notation, the scoring equations differ not only according to their type, i.e. additive or multiplicative, but also depending on their range, i.e. whether we use the signed range $[-1,+1]$ or the unsigned range $[0,1]$:

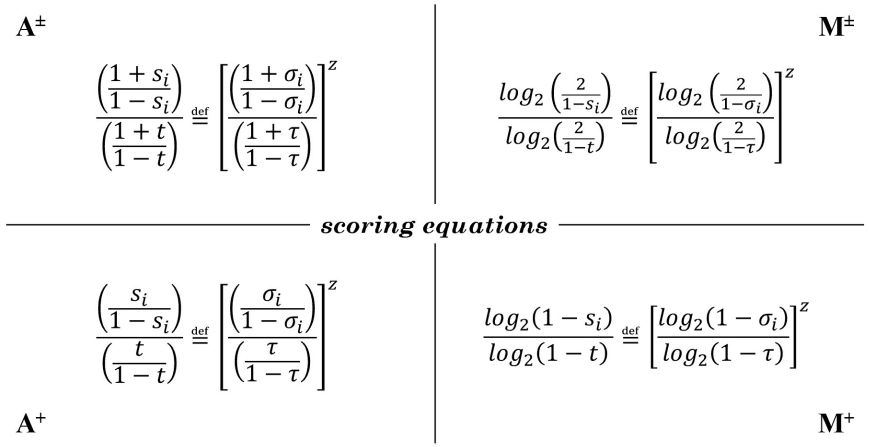

Figure 9: The scoring equations for all four scoring models in arithmetic notation (cf. Figure 2).

We end this section with some of the obvious but important properties of the scoring function which hold for all four scoring models:

(7) $f_{\tau,t}^{z}(\sigma) < f_{\tau,t}^{z}(\sigma') \quad if\ and\ only\ if \quad \sigma < \sigma'$

(8) $f_{\tau,t}^{z}(lowest\ rating) = lowest\ score \quad (either\ -1\ or\ 0)$

(9) $f_{\tau,t}^{z}(1) = 1$

(10) $f^{z}_{\tau,t}(\tau) = \mathrm{t} \quad the\ mean\ rating\ corresponds\ to\ the\ mean\ score$

(11) $f^{0}_{\tau,t}(\sigma) = \mathrm{t} \quad if\ z = 0\ then\ all\ scores\ equal\ the\ team\ score$

(12) $f^{\infty}_{\tau,t}(\sigma) = \text{lowest score} \quad if\ \sigma < \tau$

(13) $f^{\infty}_{\tau,t}(\sigma) = 1 \quad if\ \tau < \sigma$

5 The four standard TPA scoring models

In the preceding section, we have boldly postulated four basic quasi-arithmetic operations for the scoring models presented in section 1 without saying exactly how these operations are defined. Instead, we just used those to-be-defined operations to formulate three required constructs: (1) the common concept of the quasi-arithmetic mean for all scales, (2) the additive and multiplicative scoring formulae, and (3) the resulting scoring equations for the four scoring models.

In this section, we will show, for each of the four scoring models, how to define the required quasi-operations. It will turn out, that all we have to do is to specify the correct *rescaling function* for each of the four models. A rescaling function, say φ, maps the scores or ratings *either* to the real numbers (in the additive case) or to the non-negative real numbers (in the multiplicative case). The quasi-operations will then be defined in such way that they correspond in a nice and simple way to the usual arithmetic operations (so-called *Cauchy equations*, cf. Ng 2016). Here is how, where *x* and *y* are scores or ratings, and *r* is an arbitrary non-negative number (technically called scalar):

(14) $$\varphi(x \oplus y) \stackrel{\text{def}}{=} \varphi(x) + \varphi(y)$$

(15) $$\varphi(x \ominus y) \stackrel{\text{def}}{=} \varphi(x) - \varphi(y)$$

(16) $$\varphi(r \odot x) \stackrel{\text{def}}{=} r \times \varphi(x)$$

(17) $$x \oslash y \stackrel{\text{def}}{=} \varphi(x) \div \varphi(y)$$

Now, each scoring model can be characterized by a unique rescaling mapping φ. Using φ, each of the formulae given in the preceding section can be "translated" in ordinary arithmetic, as we have already done for the scoring equations (Figure 9). As an intermediary step, we may as well get rid of the special operator symbols we have introduced in the previous section. For instance, applying the rescaling func-

tion φ to the definition of the team score in equation (4), we get the usual definition of a quasi-arithmetic mean (cf. Aczél 1966):

$$(18) \quad \varphi(t) = \varphi(\bigoplus_{i=1}^{n} w_i \odot s_i) = \sum_{i=1}^{n} \varphi(w_i \odot s_i) = \sum_{i=1}^{n} w_i \times \varphi(s_i)$$

which is most often rendered in the following form (applying φ^{-1} to both sides):

$$(19) \quad t = \varphi^{-1}(\sum_{i=1}^{n} w_i \times \varphi(s_i))$$

In words: the quasi-mean of scores s_i is just like the usual arithmetic mean of those s_i, except that *before weighting* those scores are rescaled by φ and *after summing* the weighted rescaled scores the result is mapped back to the original (signed or unsigned) scale by applying the inverse of φ (the rescaling function will always be one-to-one, so that an inverse exists). It is simple to prove, that for instance the well-known geometric mean, harmonic mean and power means are just quasi-arithmetic means for some elementary φ (Jäger 2005).

Now, this is all one needs to know to understand how the four scoring models have been constructed. Of course, the crucial and difficult step is to find suitable rescaling functions φ, a different one for each of the four models. Luckily, in terms of φ, there are close relationships between the two additive models and the two multiplicative models, respectively. On the other hand, there are clear differences between the additive and the multiplicative scales which deserve fuller attention in future publications.

5.1 The signed additive scoring model $A^{\pm}$

The signed additive scoring function will be defined on the standard scale $[-1, +1]$. In practice, such scales may run from $-$n to +n, for any integer n from 1 upward, and scores or ratings on such a (2n+1)-point scale will be standardized before any further calculations are made.

To define the three operations of addition $\oplus$, subtraction $\ominus$ and scalar multiplication $\odot$ we need the following rescaling function φ: $[-1, +1] \to \mathcal{R}$ and its inverse:

$$(20) \quad \varphi(x) \stackrel{\text{def}}{=} \log\left(\frac{1+x}{1-x}\right) \qquad \varphi^{-1}(y) \stackrel{\text{def}}{=} \frac{e^y - 1}{e^y + 1}$$

Applying the Cauchy equations (14-16) we get the following definitions and properties:

Model A±	**Operations**	**Identity elements**
Addition	$x \oplus y \stackrel{\text{def}}{=} \frac{x+y}{1+x\cdot y}$	$x \oplus 0 = x$
Subtraction	$x \ominus y \stackrel{\text{def}}{=} \frac{x-y}{1-x\cdot y}$	$x \ominus 0 = x$
Scalar Multiplication	$r \odot x \stackrel{\text{def}}{=} \frac{(1+x)^r-(1-x)^r}{(1+x)^r+(1-x)^r}$	$1 \odot x = x$

Figure 10: The three operations on the scale $[-1,+1]$ for the additive model.

It is not difficult to check that these operations are well-defined, i.e. addition, subtraction and scalar multiplication (with a non-negative number) always yield valid scores or ratings, with 0 and 1 playing special roles. The formula for the quasi-arithmetic mean scoring t can now be explicitly given (the quasi-mean rating τ is defined analogously):

$$
(21)\quad \begin{cases}
t & \stackrel{\text{def}}{=} \varphi^{-1}\left(\sum_{i=1}^{n} w_i \times \varphi(s_i)\right) & by\ (19)\\
& = \varphi^{-1}\left(\sum_{i=1}^{n} w_i \times log\left(\frac{1+s_i}{1-s_i}\right)\right) & by\ (20)\\
& = \varphi^{-1}\left(\sum_{i=1}^{n} log\left(\frac{1+s_i}{1-s_i}\right)^{w_i}\right) & basic\\
& = \varphi^{-1}\left(log\prod_{i=1}^{n}\left(\frac{1+s_i}{1-s_i}\right)^{w_i}\right) & basic\\
& = \frac{e^{log\prod_{i=1}^{n}\left(\frac{1+s_i}{1-s_i}\right)^{w_i}}-1}{e^{log\prod_{i=1}^{n}\left(\frac{1+s_i}{1-s_i}\right)^{w_i}}+1} & by\ (20)\\
& = \frac{\prod_{i=1}^{n}\left(\frac{1+s_i}{1-s_i}\right)^{w_i}-1}{\prod_{i=1}^{n}\left(\frac{1+s_i}{1-s_i}\right)^{w_i}+1} & basic\\
& = \frac{\prod_{i=1}^{n}(1+s_i)^{w_i}-\prod_{i=1}^{n}(1-s_i)^{w_i}}{\prod_{i=1}^{n}(1+s_i)^{w_i}+\prod_{i=1}^{n}(1-s_i)^{w_i}} & basic
\end{cases}
$$

Taking up the scoring function for the additive model on $[-1,+1]$ from formula (6) and applying the foregoing definitions we get:

$$
(22)\begin{cases}
s & \stackrel{\text{def}}{=} t \oplus [z \odot (\sigma \ominus \tau)] & by\ (6)\\
& = \varphi^{-1}\left(\varphi(t) + z \times \left(\varphi(\sigma) - \varphi(\tau)\right)\right) & by\ (14-16)\\
& = \varphi^{-1}\left(log\left(\frac{1+t}{1-t}\right) + z \times \left(log\left(\frac{1+\sigma}{1-\sigma}\right) - log\left(\frac{1+\tau}{1-\tau}\right)\right)\right) & by\ (20)\\
& = \varphi^{-1}\left(log\left[\left(\frac{1+t}{1-t}\right) \times \left(\frac{1+\sigma}{1-\sigma}\right)^{z} / \left(\frac{1+\tau}{1-\tau}\right)^{z}\right]\right) & basic\\
& = \frac{\left(\frac{1+t}{1-t}\right)\times\left(\frac{1+\sigma}{1-\sigma}\right)^{z} / \left(\frac{1+\tau}{1-\tau}\right)^{z} - 1}{\left(\frac{1+t}{1-t}\right)\times\left(\frac{1+\sigma}{1-\sigma}\right)^{z} / \left(\frac{1+\tau}{1-\tau}\right)^{z} + 1} & by\ (20)\\
& = \frac{\left(\frac{1+\sigma}{1-\sigma}\right)^{z}(1+t) - \left(\frac{1+\tau}{1-\tau}\right)^{z}(1-t)}{\left(\frac{1+\sigma}{1-\sigma}\right)^{z}(1+t) + \left(\frac{1+\tau}{1-\tau}\right)^{z}(1-t)} & basic
\end{cases}
$$

5.2 The unsigned additive scoring model A^+

The unsigned additive scoring function will be defined on the standard scale $[0,1]$. In practice, such scales may run from 1 to n, for any integer n from 2 upward, and scores or ratings on such a n-point scale will be standardized before any further calculations are made.

To define the three operations of addition, subtraction and scalar multiplication we need the following rescaling function and its inverse:

$$
(23) \qquad \varphi(x) \stackrel{\text{def}}{=} \log\left(\frac{x}{1-x}\right) \qquad \varphi^{-1}(y) \stackrel{\text{def}}{=} \frac{e^{y}}{e^{y}+1}
$$

Applying the Cauchy equations (14-16) we have the following definitions:

Model A^+	**Operations**	**Identity elements**
Addition	$x \oplus y \stackrel{\text{def}}{=} \frac{x \cdot y}{x \cdot y \ + (1-x) \cdot (1-y)}$	$x \oplus ½ = x$
Subtraction	$x \ominus y \stackrel{\text{def}}{=} \frac{x \div y}{x \div y \ + (1-x) \div (1-y)}$	$x \ominus ½ = x$
Scalar Multiplication	$r \odot x \stackrel{\text{def}}{=} \frac{x^{r}}{x^{r} + (1-x)^{r}}$	$1 \odot x = x$

Figure 11: The three operations on the scale [0,1] for the additive model.

It is not difficult to check that these operations are well-defined, i.e. addition, subtraction and scalar multiplication (with a non-negative number) always yield valid

scores or ratings; ½ and 1 are the neutral elements. The formula for the quasi-arithmetic mean scoring t can now be explicitly given (the quasi-mean rating τ is defined analogously):

$$
(24)\quad \begin{cases} t & \stackrel{\text{def}}{=} \varphi^{-1}\left(\sum_{i=1}^{n} w_i \times \varphi(s_i)\right) & by\ (19) \\ & = \varphi^{-1}\left(\sum_{i=1}^{n} w_i \times log\left(\frac{s_i}{1-s_i}\right)\right) & by\ (23) \\ & = \varphi^{-1}\left(\sum_{i=1}^{n} log\left(\frac{s_i}{1-s_i}\right)^{w_i}\right) & basic \\ & = \varphi^{-1}\left(log \prod_{i=1}^{n}\left(\frac{s_i}{1-s_i}\right)^{w_i}\right) & basic \\ & = \frac{e^{log \prod_{i=1}^{n}\left(\frac{s_i}{1-s_i}\right)^{w_i}}}{e^{log \prod_{i=1}^{n}\left(\frac{s_i}{1-s_i}\right)^{w_i}} + 1} & by\ (23) \\ & = \frac{\prod_{i=1}^{n}\left(\frac{s_i}{1-s_i}\right)^{w_i}}{\prod_{i=1}^{n}\left(\frac{s_i}{1-s_i}\right)^{w_i} + 1} & basic \\ & = \frac{\prod_{i=1}^{n}(s_i)^{w_i}}{\prod_{i=1}^{n}(s_i)^{w_i} + \prod_{i=1}^{n}(1-s_i)^{w_i}} & basic \end{cases}
$$

Taking up the scoring function for the additive model on [0,1] from formula (10) and applying the foregoing definitions we get:

$$
(25)\quad \begin{cases} s & \stackrel{\text{def}}{=} t \oplus [z \odot (\sigma \ominus \tau)] & by\ (6) \\ & = \varphi^{-1}\left(\varphi(t) + z \times \left(\varphi(\sigma) - \varphi(\tau)\right)\right) & by\ (14-16) \\ & = \varphi^{-1}\left(log\left(\frac{t}{1-t}\right) + z \times \left(log\left(\frac{\sigma}{1-\sigma}\right) - log\left(\frac{\tau}{1-\tau}\right)\right)\right) & by\ (23) \\ & = \varphi^{-1}\left(log\left[\left(\frac{t}{1-t}\right) \times \left(\frac{\sigma}{1-\sigma}\right)^{z} / \left(\frac{\tau}{1-\tau}\right)^{z}\right]\right) & basic \\ & = \frac{\left(\frac{t}{1-t}\right) \times \left(\frac{\sigma}{1-\sigma}\right)^{z} / \left(\frac{\tau}{1-\tau}\right)^{z}}{\left(\frac{t}{1-t}\right) \times \left(\frac{\sigma}{1-\sigma}\right)^{z} / \left(\frac{\tau}{1-\tau}\right)^{z} + 1} & by\ (23) \\ & = \frac{\left(\frac{\sigma}{1-\sigma}\right)^{z}(t)}{\left(\frac{\sigma}{1-\sigma}\right)^{z}(t) + \left(\frac{\tau}{1-\tau}\right)^{z}(1-t)} & basic \end{cases}
$$

5.3 The signed multiplicative scoring model $M^{\pm}$

The signed multiplicative scoring function will be defined on the standard scale $[-1, +1]$. In practice, such scales may run from −n to +n, for any integer n from 1 upward, and scores or ratings on such a (2n+1)-point scale will be standardized before any further calculations are made.

To define the operations of addition $\oplus$, scalar multiplication $\odot$ and scalar division $\oslash$ we need another rescaling function φ^{-1}: $[-1, +1] \to \mathcal{R}$ and its inverse:

$$(26)\qquad \begin{cases} \varphi(x) & \stackrel{\text{def}}{=} 1-\log_2(1-x)=\log_2\left(\frac{2}{1-x}\right) \\ \varphi^{-1}(y) & \stackrel{\text{def}}{=} 1-2^{1-y}=1-2\cdot 2^{-y} \end{cases}$$

Applying the Cauchy equations (14, 16-17) we get the following definitions:

Model M±	**Operations**	**Identity elements**
Addition	$x \oplus y \stackrel{\text{def}}{=} 1-2\cdot\left(\frac{1-x}{2}\right)\cdot\left(\frac{1-y}{2}\right)$	$x \oplus 0 = x$
Scalar Multiplication	$r \odot x \stackrel{\text{def}}{=} 1-2\cdot\left(\frac{1-x}{2}\right)^r$	$1 \odot x = x$
Scalar Division	$x \oslash y \stackrel{\text{def}}{=} log_2\left(\frac{2}{1-x}\right) \div log_2\left(\frac{2}{1-y}\right)$	$x \oslash 0 = \varphi(x)$

Figure 12: The three operations on the scale [−1,+1] for the multiplicative model.

It is not difficult to check that these operations are well-defined, i.e. addition and scalar multiplication (with a non-negative number) always yield valid scores or ratings and scalar division always yields a scalar; 0 and 1 are the neutral elements. The formula for the quasi-arithmetic mean scoring t can now be explicitly given (the quasi-mean rating τ is defined analogously):

$$(27)\qquad \begin{cases} t & \stackrel{\text{def}}{=} \varphi^{-1}\left(\sum_{i=1}^n w_i \times \varphi(s_i)\right) & by\ (19) \\ & = \varphi^{-1}\left(\sum_{i=1}^n w_i \times (1-\log_2(1-s_i))\right) & by\ (26) \\ & = \varphi^{-1}\left(1-\sum_{i=1}^n w_i \times \log_2(1-s_i)\right) & basic \\ & = 1-2^{\sum_{i=1}^n w_i \times \log_2(1-s_i)} & by\ (26) \\ & = 1-\prod_{i=1}^n 2^{w_i \times \log_2(1-s_i)} & basic \\ & = 1-\prod_{i=1}^n (1-s_i)^{w_i} & basic \end{cases}$$

Taking up the scoring function for the signed multiplicative model on [−1,+1] from formula (6) and applying the foregoing definitions we get:

$$
(28)\quad \begin{cases} s & \stackrel{\text{def}}{=} (\sigma \oslash \tau)^z \odot t & by\ (6) \\ & = \varphi^{-1}\left(\left(\frac{\varphi(\sigma)}{\varphi(\tau)}\right)^z \times \varphi(t)\right) & by\ (14-17) \\ & = \varphi^{-1}\left(\left(\frac{\log_2\left(\frac{2}{1-\sigma}\right)}{\log_2\left(\frac{2}{1-\tau}\right)}\right)^z \times \log_2\left(\frac{2}{1-t}\right)\right) & by\ (26) \\ & = 1 - 2 \cdot 2^{-\left(\frac{\log_2\left(\frac{2}{1-\sigma}\right)}{\log_2\left(\frac{2}{1-\tau}\right)}\right)^z \times \log_2\left(\frac{2}{1-t}\right)} & by\ (26) \\ & = 1 - 2 \cdot \left(\frac{1-t}{2}\right)^{\left(\frac{\log_2\left(\frac{2}{1-\sigma}\right)}{\log_2\left(\frac{2}{1-\tau}\right)}\right)^z} & basic \end{cases}
$$

5.4 The unsigned multiplicative scoring model M⁺

The unsigned multiplicative scoring function will be defined on the standard scale [0,1]. In practice, such scales may run from 1 to n, for any integer n from 2 upward, and scores or ratings on such a n-point scale will be standardized before any further calculations are made.

$$
(29)\qquad \begin{cases} \varphi(x) & \stackrel{\text{def}}{=} \log_2\left(\frac{1}{1-x}\right) = -\log_2(1-x) \\ \varphi^{-1}(y) & \stackrel{\text{def}}{=} 1 - 2^{-y} \end{cases}
$$

Applying the Cauchy equations (14, 16-17) we get the following definitions:

Model M⁺	**Operations**	**Identity elements**
Addition	$x \oplus y \stackrel{\text{def}}{=} 1 - (1-x) \cdot (1-y)$	$x \oplus ½ = x$
Scalar Multiplication	$r \odot x \stackrel{\text{def}}{=} 1 - (1-x)^r$	$1 \odot x = x$
Scalar Division	$x \oslash y \stackrel{\text{def}}{=} log_2(1-x) \div log_2(1-y)$	$x \oslash ½ = \varphi(x)$

Figure 13: The three operations on the scale [0,1] for the multiplicative model.

It is not difficult to check that these operations are well-defined, i.e. addition and scalar multiplication (with a non-negative number) always yield valid scores or ratings and scalar division always yields a scalar; ½ and 1 are neutral elements. The formula for the quasi-arithmetic mean scoring t can now be explicitly given (the quasi-mean rating τ is defined analogously):

$$(30)\quad \begin{cases} t & \stackrel{\text{def}}{=} \varphi^{-1}(\sum_{i=1}^{n} w_i \times \varphi(s_i)) & by\ (19) \\ & = \varphi^{-1}(\sum_{i=1}^{n} w_i \times (-\log_2(1-s_i))) & by\ (29) \\ & = \varphi^{-1}(-\sum_{i=1}^{n} w_i \times \log_2(1-s_i)) & basic \\ & = 1 - 2^{\sum_{i=1}^{n} w_i \times \log_2(1-s_i)} & by\ (29) \\ & = 1 - \prod_{i=1}^{n} 2^{w_i \times \log_2(1-s_i)} & basic \\ & = 1 - \prod_{i=1}^{n} (1-s_i)^{w_i} & basic \end{cases}$$

Taking up the scoring function for the additive model on [0,1] from formula (10) and applying the foregoing definitions we get:

$$(31)\quad \begin{cases} s & \stackrel{\text{def}}{=} (\sigma \oslash \tau)^z \odot t & by\ (6) \\ & = \varphi^{-1}\left(\left(\frac{\varphi(\sigma)}{\varphi(\tau)}\right)^z \times \varphi(t)\right) & by\ (14-17) \\ & = \varphi^{-1}\left(-\left(\frac{\log_2(1-\sigma)}{\log_2(1-\tau)}\right)^z \times \log_2(1-t)\right) & by\ (29) \\ & = 1 - 2^{\left(\frac{\log_2(1-\sigma)}{\log_2(1-\tau)}\right)^z \times \log_2(1-t)} & by\ (29) \\ & = 1 - (1-t)^{\left(\frac{\log_2(1-\sigma)}{\log_2(1-\tau)}\right)^z} & basic \end{cases}$$

6 From Excel to professional software tools

Currently, our TPA models are mainly implemented and applied using Excel, with or without the help of VBA (Visual Basic for Applications). Here is an example.

The example consists of two so-called dashboards. The first dashboard (Figure 14) is mainly used for recording the names of the students ("peers") in the team and their mutual peer ratings. It consists of the following parts:

- A: Lecturer's name or ID
- B: Course name or ID
- C: Team name or ID
- D: Name or ID's of the peers + student weights (in percentages)
- E: Peer to be assessed (drop-down menu)
- F: Peer ratings from the other students (slider)
- G: Peer rating matrix (lower part) + Student ratings (upper part)

Figure 14: Implementation of a TPA model in Excel-VBA: Dashboard I.

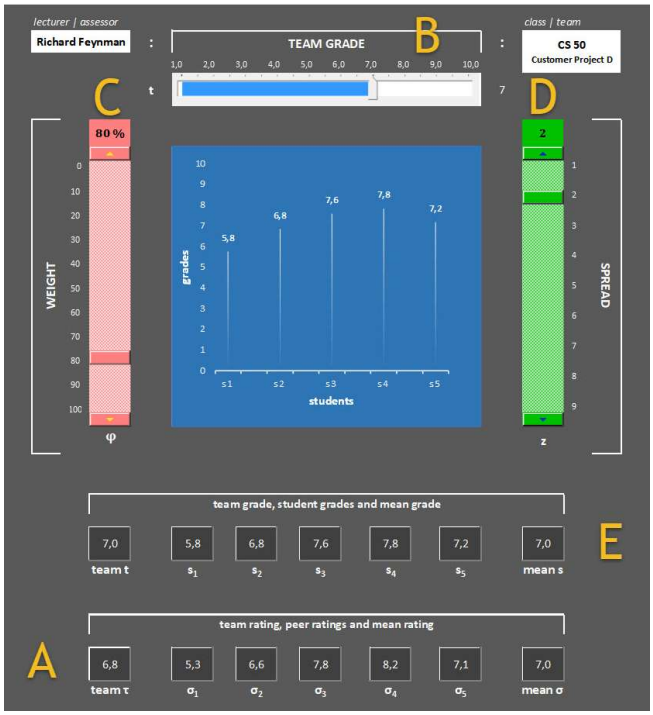

Figure 15: Implementation of a TPA model in Excel-VBA: Dashboard II.

The second dashboard (Figure 15) is used by the lecturer to enter the team score, the team weighting (a parameter not used any more) and the zoom factor, and for calculation and graphical presentation of the final student scores. It consists of the following parts:

- A: Replication of the student ratings (taken from Dashboard I)
- B: A slider for entering the team score (here dubbed *grade*)
- C: A slider for entering a team weighting factor (*not used any more)
- D: A slider for entering the zooming factor
- E: Presentation of the final student scores

Our long-term goal is to develop and distribute a fully stand-alone software package written in one of the mainstream programming languages, e.g. Java, with an excellent User-Friendly Interface. As a first step in this direction we have already run two student projects, one rather small end-of-course assignment, the other a more challenging bachelor thesis project, in which selected modules of the Team-Peer-Assessment system have been programmed (Cresta 2018).

In general, the latter prototype follows the logic of the Excel-VBA implementation shown above but it permits *on-the-fly* selection of one of the scoring models and it has a more modern look-and-feel. The results are very encouraging, but we need more of such pilot projects to convince commercial software developers of the potential merits of our TPA approach.

7 Conclusion

In this paper we have presented a detailed account of our innovative Team-Peer-Assessment approach based on an advanced theory of scoring and rating based on fuzzy algebra. For the first time we have shown that it is possible to consistently define so-called additive and multiplicative scoring models for peer assessment. These scoring models are based on a few sound principles and behave correctly and fairly, primarily thanks to the Split-Join-Invariance principle.

It is now up to the community of educational practitioners and researchers to investigate the suitability and practicality of the approach in the classroom. Many details and advantages of our approach will only become clear when practitioners and researchers compare their own approach with our TPA proposal.

We will be glad to offer support to initiate such pilot projects in the context of teamwork-based courses.

8 Glossary

The definitions pertain specifically to this paper, if not stated otherwise.

Arithmetic mean (AM): A well-known quantity used to represent a set of measures of one and the same phenomenon. Often used in the context of statistical data analysis, but that is not a requirement. Because the measures are added together after weighting them by weighting factors (all summing to one) to arrive at the quantity, it assumes that the measures can take on any real value. If measures are restricted to bounded ranges (intervals), arithmetic means may give distorted results.

Differential scoring: In the context of teamwork as an assessment method, a type of scoring that may produce different scores for different students – in contrast to the default approach by which all students will get the same (team) score.

Geometric mean (GM): Another well-known quantity used to represent a set of measures of one and the same phenomenon. Often used in the context of statistical data analysis, but not a requirement. Because the measures are multiplied with each other after raising them to a exponential weight (all summing to one) to arrive at the quantity, it assumes that the measures can take on only non-negative real values. If measures are restricted to other bounded ranges (intervals), geometric means may give distorted results.

Mean: Generally, a quantity assumed to faithfully represent a set of given measures of one and the same phenomenon. Often used in the context of statistical data analysis. Here, we use the algebraic notion of mean, which is precisely defined by a small number of axioms. See e.g. Kolmogorov 1930.

Mean score (t): In the context of teamwork, the mean score is the score which results from calculating the mean of the individual scores of all members of that team.

Mean team rating (τ): In the context of team assessment, the mean team rating is the **rating** which results from calculating the mean of the individual ratings of the students in a team. If there can be no confusion, it will also be called mean rating or team rating.

Mean peer rating (σ_i): In the context of **TPA**, the mean peer rating is the **rating** which results from calculating a mean (**QAM** or **QGM**) of all **peer ratings** for one student (with index i) in that team.

Peer assessment: Part of **TPA**'s procedure in which the students of a team assess each other. Another part of TPA is the criteria-based assessment of a team's product which is conducted by the lecturer to set the **team score**. Finally, the set of formulae **(**mainly **quasi-arithmetic mean** and **scoring formula)**, which glue all the different measurements together may be thought of as a third part of **TPA**.

Peer matrix: A square matrix of order n (the size of the team) which systematically captures all peer ratings of all members of a team. Self-ratings (a student rating him- or herself) are not foreseen in TPA.

Peer rating (σ_{ij}): In the context of **TPA**, a peer rating is the rating which results from calculating the mean (**QAM** or **QGM**) of all criteria-based judgments from one student about another peer in that team.

Quasi-addition ($\oplus$): A quasi-addition is like common arithmetical addition on the reals with the difference that the range of admissible values is restricted to a subset of the reals. Here, we only consider the two standard ranges **signed unit interval** and **unsigned unit interval**, as usual in fuzzy mathematics. Furthermore, quasi-addition requires a so-called rescaling function (here always denoted by ϕ) which uniquely defines the rescaled quasi-sum of x and y as the sum of the rescaled values of x and y.

Quasi-subtraction ($\ominus$): A quasi-subtraction is like common arithmetical subtraction on the reals with the difference that the range of admissible values is restricted to a subset of the reals. Here, we only consider the two standard ranges **signed unit interval** and **unsigned unit interval**, as usual in fuzzy mathematics. Furthermore, quasi-subtraction requires a so-called rescaling function (here always denoted by ϕ) which uniquely defines the rescaled quasi-subtraction of x and y as the difference between the rescaled values of x and y, in that order.

Quasi-multiplication ($\odot$): A quasi-multiplication is like common scalar multiplication of x with a non-negative real number r (called scalar) with the only difference that the range of admissible x is restricted to a subset of the reals. Here, we only consider the two standard ranges **signed unit interval** and **unsigned unit interval**, as usual in fuzzy mathematics. Furthermore, quasi-multiplication requires a so-called rescaling function (here always denoted by ϕ) which uniquely defines the rescaled quasi-product of x with the scalar r as the product of r and the rescaled value of x.

Quasi-division ($\oslash$): A quasi-division is like common arithmetical division on the reals with the difference that the range of admissible values is restricted to a subset of the reals. Here, we only consider the two standard ranges **signed unit interval** and **unsigned unit interval**, as usual in fuzzy mathematics. Furthermore, quasi-division requires a so-called rescaling function (here always denoted by ϕ) which uniquely defines the rescaled quasi-quotient of x and y as the quotient of the rescaled values of x and y, which is a scalar, not a score or rating.

Quasi-arithmetic mean (QAM): A quasi-arithmetic mean is like the usual arithmetic mean with one important difference which makes it much more flexible: the mean is not calculated on the basis of the raw measurements; instead a rescaling of those raw measurements is performed before calculating the mean, and afterwards the resulting mean is send back to the original scale by applying the inverse of the rescaling function. It can be proven that many well-known means, e.g. the geometric mean, are just quasi-arithmetic means for some rescaling function.

Quasi-geometric mean (QGM): A quasi-geometric mean is like the usual geometric mean with one important difference: the mean is not calculated on the basis of the given measurements; instead a rescaling of those raw measurements is performed before calculating the geometric mean, and afterwards the resulting mean is scaled back to the original scale by applying the inverse of the rescaling

function. It can be proven that many well-known means (geometric mean, harmonic mean, power mean, etc.) are just quasi-arithmetic means for some appropriate rescaling function

Rating (σ): Rating is a form of human measurement, in a double sense. Firstly, it means the judgement of human performance, attitudes or characteristics. Secondly, it means the measurement conducted by human beings in contrast to some physical measurement procedure or equipment. Here, in the context of **peer assessment**, we have adopted the term rating for the mutual judgment of students (peers) in a team on diverse process quality criteria.

Rating scale: A range (interval) of values on the real line chosen for **rating** a well-defined attitude, characteristic or performance of a person or group of persons. As **ratings** are (usually) judged by human beings (judges) these measurements (judgments) may be subject to well-known forms of bias and should therefore be handled with care. A way to prevent some forms of bias is to provide adequate training of the judges and/or to collect multiple ratings of the same factor by a team of raters.

Scale: A range (interval) of values on the real line chosen for measuring a given characteristic or performance of a person or group of persons. Values which don't belong to the chosen range aren't admissible as valid measurements of the characteristic or performance being measured. Sometimes, a scale is defined as such a range together with some well-defined relations and operations on the numbers in that range (e.g. equality, addition, multiplication, etc.).

Scoring (s): Scoring is a form of human measurement, in a double sense. Firstly, it means the judgement of human performance, attitudes or characteristics. Secondly, it means the measurement conducted by human beings - in contrast to some physical measurement procedure or equipment. Here, in the context of **TPA**, we have adopted the term scoring for the judgment of a team's products (output, deliverables) by the lecturer and the subsequent calculation of individual student scores based on the adopted scoring rule (**scoring function**) on diverse process quality criteria.

Scoring function ($f_{\tau,t}^{z}$): A mapping from the **rating scale** to the **scoring scale** which takes a given **student rating** and maps it on the corresponding **student score**. Used to differentiate (moderate) between students who have been working differently in the same team. The scoring function requires three parameters: the **team score** t as set by the lecturer, the **team rating** τ calculated from all individual **student ratings**, and the **zooming parameter** z.

Scoring scale: The measurement scale used for scoring the work of students or teams of students in educational assessment. Here, two standard scoring scales are used (cf. **signed unit interval** or **unsigned unit interval**).

Signed unit interval ([0, +1]): All real numbers between (inclusive) 0 and +1.

Student rating (σ or σ_i): In the context of **TPA**, the student rating is the **rating** which results from calculating a mean (**QAM** or **QGM**) of all **peer ratings** for a single student in the team.

Student score (s_i): A student score is the final score for work done on an educational assignment, in teamwork or not. Here we only consider teamwork, so that

a student score is a student's individual score resulting from the teamwork. This score usually encompasses the assessment of both product and process quality criteria. In the context of TPA, product criteria are assessed by the lecturer, and process criteria are assessed by the students (**peer assessment**).

Team rating: see Mean team rating

Team score (t): The mean (**QAM** or **QGM**) of all individual student scores.

Team-Peer-Assessment (TPA): A well-founded framework for educational assessment, which enables lecturers to assign **differential scores** to students who have been working on a team assignment. It assumes, that lecturers are best equipped to judge the quality of the team product (what they deliver), while students are best equipped to judge the quality of the team process (how they deliver). Based on a few principles, the most important of which are: using **quasi-arithmetic means** on all levels of aggregation, and the **split-join-invariance** principle.

Split-Join-Invariance (SJI): Basic principle of **TPA**. To calculate the individual score of a student, a scoring formula must be applied, which has the **team score** as one of its three parameters. This team score will be calculated based on the team's products as the mean performance on a set of product quality criteria judged by the lecturer (his unique competence, see **TPA**). Once the individual **student scores** have been calculated, one can again calculate the **team score** using the adopted **QAM** or **QGM**. The SJI now forces and guarantees that this calculated team score is exactly equal to the initial team score provided by the lecturer.

Unsigned unit interval ([-1, +1]): The real numbers between (inclusive) -1 and +1.

Weights (w or w_i): In the context of calculating means, weights are real numbers between 0 and 1 (inclusive) which are used as multipliers (**AM** or **QAM**) or exponents (**GM**, **QGM**) in the formula for the mean. The sum of all weights should be 1. Note that weights have nothing to do with probability!

Zoom(ing) factor (z): To adjust (moderate) the impact of student ratings on lecturer's score (equal to the **team score**) in the scoring formula, a zooming factor or parameter can be used. A zooming factor is a non-negative real number. Its default value in all scoring formulae is 1. Choosing a zooming factor smaller than 1 decreases the impact of students' ratings on the team score; in the extreme case, $z = 0$, there will be no adjusting or moderating at all. Choosing a zooming factor larger than 1 increases the impact of students' ratings on the team score; in the extreme, the student score will become as low as possible on the scale (-1 or 0) or as high as possible (always 1).

References

Aczél, J. (Ed.) (1966). Lectures on functional equations and their applications (Vol. 19). Academic Press.

Batchelder, W. H., Colonius, H., Dzhafarov, E. N., & Myung, J. (Eds.) (2016). New Handbook of Mathematical Psychology: Volume 1, Foundations and Methodology. Cambridge University Press

Bukowski, W. M., Castellanos, M., & Persram, R. J. (2017). The current status of peer assessment techniques and sociometric methods. In Peter E. L. Marks & Antonius H. N. Cillessen (Eds.), New Directions in Peer Nomination Methodology. New Directions for Child and Adolescent Development, 157, 75–82.

Cresta, R.-A. (2018). Peer Performance Scoring System (PPASS). Technical report, BSc Computing (Software Engineering) dissertation, University of Northampton, 148p.

Dijkstra, J., Latijnhouwers, M., Norbart, A., & Tio, R. A. (2016). Assessing the "I" in group work assessment: State of the art and recommendations for practice. *Medical Teacher*, *38*(7), 675-682.

Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher education*, *24*(3), 331-350.

Dombi, J. (1982). A general class of fuzzy operators, the DeMorgan class of fuzzy operators and fuzziness measures induced by fuzzy operators. Fuzzy sets and systems, 8(2), 149-163.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self-assessments. Assessment and Evaluation in Higher Education, 11(2), 146-165.

Falchikov, N. (1993). Group process analysis: self and peer assessment of working together in a group. Educational and Training Technology, 30(3), 275–284.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research*, *70*(3), 287-322.

Fodor, J. (2009). Aggregation Functions in Fuzzy Systems. In: J. Fodor & J. Kacprzyk (Eds.): Aspects of Soft Computing, Intelligent Robotics & Control, SCI 241, 25–50.

Gibbs, G. (2009). The assessment of group work: lessons from the literature. Assessment Standards Knowledge Exchange.

Jäger, J. (2005). Verknüpfungsmittelwerte, Math. Semesterberichte, 52: 63-80.

Kennedy, I. G., & Vossen, P. H. (2017a). Teamwork assessment and peerwise scoring: Combining process and product assessment. DeLFI, Leipzig: Bildungsräume, Lecture Notes in Informatics, Gesellschaft für Informatik, Bonn

Kennedy, I. G., & Vossen, P. H. (2017b). Software engineering teamwork assessment rubrics: combining process and product scoring. Assessment in Higher Education, Manchester

Kolmogorov, A. (1930) "On the Notion of Mean", in "Mathematics and Mechanics" (Kluwer 1991), 144–146.

Nepal, K. P. (2012). An approach to assign individual marks from a team mark: the case of Australian grading system at universities. *Assessment & Evaluation in Higher Education*, *37*(5), 555-562.

Ng, C.T., (2016) Functional equations. In: Batchelder, W. H., Colonius, H., Dzhafarov, E. N., & Myung, J. (Eds.). (2016). New Handbook of Mathematical Psychology: Volume 1, Foundations and Methodology. Cambridge University Press, 151-193

Sharp, S. (2006). Deriving individual student marks from a tutor's assessment of group work. Assessment & Evaluation in Higher Education, 31(3), 329-343.

Strijbos, J. W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction*, *20*(4), 265-269.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of educational Research*, *68*(3), 249-276.

Ueno, M. (2010). An Item Response Theory for Peer Assessment, In: Mary Beth Rosson (Ed.), Advances in Learning Processes, InTech

Uto, M., & Ueno, M. (2016). Item Response Theory for Peer Assessment. *IEEE transactions on learning technologies*, *9*(2), 157-170.

Van Rensburg, J. (2012). Assessment of Group Work: Summary of a Literature Review.

Vossen, P.H. (2018). Distributive Fairness in Educational Assessment: Psychometric Theory meets Fuzzy Logic. In: Balas et al. (eds.), Soft Computing Applications, Advances in Intelligent Systems and Computing 634, Springer International Publishing, p. 381-394