

# Comproved: Why make assessing difficult when it can be easy?



*ECHO Tip July 2021*

---

*Centre of Expertise for Higher Education (University of Antwerp)*

---

*In collaboration with Maarten Goossens (Training and Education Sciences, and co-founder of Comproved, UAntwerp), Karla Groen (E-campus, UAntwerp) and Philip Lambrechts (Department of Education, UAntwerp)*

Lecturers often use **criteria lists**, called [rubrics](#) (info in Dutch), to assess student performance. This way, they can determine the extent to which students have achieved certain competences in an **objective manner**. However, using criteria or rubrics to assess students isn't always self-evident. Do any of the following situations sound familiar? (1) A fellow lecturer has drawn up a clever, detailed rubric for assessing writing assignments. Great, so now both of you have the same frame of reference for awarding marks. However, as you're marking assignments, you often pause and wonder what exactly a certain competence description means. (2) You read through a student's completed assignment, and based on your experience, you estimate that it deserves 12/20. However, after adding up the marks for the predefined criteria, you end up with a total score of 15/20. A significant difference! Do you stick to the objective criteria and give the student a 15, or do you go back and tweak some marks to arrive at a total score of 12? (3) You give students an extensive assignment that's meant, in part, to stimulate and assess their creativity. How do you define a complex competence like creativity so that everyone understands the definition in the same way, but without losing sight of certain nuances and of the overall picture? Can all aspects of creativity really be captured in a set of predefined criteria (see also [the ECHO teaching tip](#) (info in Dutch) on stimulating and assessing creativity)?

It's not that strange if you've ever found yourself in (one of) these situations. After all, scientific research has shown that there are some **pitfalls** when **developing and using criteria** and **rubrics** (Bloxham, 2009; Sadler, 2009). The **pairwise comparison method** provides opportunities to overcome these difficulties. In what follows, we'll start by clarifying the concept of pairwise comparison. Then we'll look at some specific applications using the Comproved tool. We'll conclude this tip with an overview of some key aspects to keep in mind when using pairwise comparison and/or Comproved.

---

## Pairwise comparison

---

Laming (2014) states that 'every judgment is a comparison'. For example, when assessing a student's work, you compare it either to other students' work or to your internal standard. Even if you use objective standards, like a set of criteria, you're still comparing – either to that standard or to other tasks you've judged against that standard. In other words, **comparing is an implicit part of assessing**. Pairwise comparison makes this implicit comparison explicit. This method is rooted in

Thurstone's Law (1927), which states that people are better at comparing two objects and determining which one is better than they are at assigning absolute scores to a single object. Let's illustrate this with an example:

*You're handed an unfamiliar object and asked to estimate how much it weighs. This is a difficult task, which feels like taking a leap in the dark, and chances are very slim that you'll be able to guess the exact weight. Now, let's say you're handed two objects and asked which one is heavier: this seems*





*like a very easy task. You feel like this won't be any trouble at all, and you'll probably get it right.*

In the pairwise comparison method, all objects to be assessed are put into pairs **at random**, and **multiple assessors assess every pair**. In education, this means that several teachers compare pairs of tasks performed by different students and indicate **which of the two is better** when it comes to a certain competence. This is clear and simple. On the basis of the choices made, a **reliable ranking can then be made** from best to worst.

This method deliberately calls on people's expertise as assessors, without forcing them to look through a specific lens (e.g. a criteria list or a rubric). As any given assignment is assessed several times, in different pairs, by different assessors, **all aspects of the competence to be assessed are taken into account** in the final score. This ensures the **validity** of the assessment method (see the ECHO Teaching Tip '[Measure what you want to know](#)' from 2013).

### Potential for use in education

In education, pairwise comparison can be used to assess products of various kinds and from **different contexts**, including mathematical reasoning tasks, writing and design products, presentations, etc. An advantage of pairwise comparison is the **potential openness of the tasks**. As you develop an assignment, you don't have to ask yourself: '(How) can this question be marked?' For example, you don't have to judge whether someone's argumentation is right or wrong, as you can simply indicate which of the two is the strongest – even if neither yields the correct result because small calculation errors were made.

You can also use pairwise comparison as a **teaching method** in your teaching practice. For instance, you can present two sample texts and encourage your students to look for reasons why one is better than the other. As the students are allowed to compare, they often notice the different characteristics of each text. One text may have a much clearer structure than the other. And in the next pair, they may notice that the

arguments put forward in the second text make more sense than those in the first. This allows students to **develop their own quality criteria** in a simple way.

If you want to use pairwise comparison as an **assessment method**, it's best to have technical support from a specific tool. One of the tools providing such support is **Comproved**, designed by researchers from UAntwerp, UGent and Imec.

### Specific cases using Comproved

Since pairwise comparison is a relatively simple task that doesn't require extensive training or briefing, it lends itself very well to **peer feedback** and **peer assessment** (see the ECHO tip '[Reliability of peer assessment](#)' from 2017). After all, students don't need to make any **absolute statements** about the work of their fellow students, and everything is done **anonymously**. This often eliminates **reluctance** to engage in peer feedback and peer assessment and the **feelings of insecurity** that sometimes come with it.

Below are two specific descriptions of peer feedback and peer assessment with **pairwise comparison done through Comproved**.

- (1) Peer feedback on 'mood boards' in the Interior Architecture programme at UAntwerp

*A **mood board** is a collection of images that express a certain emotion, often used to communicate visually with the customer about atmospheres, feelings, organisational values, and so on.*

*Every year, Interior Architecture students at UAntwerp are given the assignment to develop mood boards as a group. These works are then **discussed** in class, so the **students can learn to gauge the quality of a work**. This is usually an **arduous and time-consuming process**.*

*In order to make this process **easier and more efficient**, a form of **peer feedback in Comproved** was developed. This allowed the students to **assess and comment** on all mood boards **at home**.*



Specifically, the students' mood boards were uploaded to Comproved and **paired up at random** by the tool's algorithm. The students then had to select the best of two mood boards and indicate the **strengths and weaknesses** of each. They had a week to complete this task. When they signed in again, they picked up where they left off, until a predetermined number of comparisons was made.

Based on all the choices made by the students, Comproved generated a **ranking** of the uploaded works, from lowest to highest quality, **as estimated by the students**.

This ranking served as a basis for the lecturer to discuss the mood boards in class. This was done by **projecting and discussing the results from Comproved in a feedback lecture**. Both the ranking itself (based on the students' versus the lecturer's qualitative assessments) and the feedback given (what aspects did students pay attention to and what was their reasoning) were used for this.

Afterwards, the **feedback was sent to the respective groups** so they could use it to adjust their mood boards.

This pass/fail assessment turned out to be **in line with the peer assessment in Comproved**. Out of a hundred papers, fourteen got a failing mark from the lecturer, and these fourteen were among the twenty lowest-ranked papers in the ranking that resulted from the student comparisons in Comproved. In other words, those papers were **also perceived by the students as lower in quality**. Moreover, the **feedback** provided by the students on the papers turned out to be **very useful** and comparable to the feedback provided by the lecturer.

Based on these results, the lecturer decided that from then on, papers would only be assessed and receive feedback through peer assessment. The lecturer only checked the **bottom forty per cent** of the ranking to ensure there weren't **any unjustified 'fails'** and to determine the pass mark. This method **saved the lecturer a considerable amount of time**.

---

## Recap

---

### How does comparative assessment work?

- Each **assessor** (lecturer, external assessor, student) compares **several pairs of assignments**.
- **Multiple assessors** make these comparisons, so that every assignment is compared several times.
- On this basis, a **ranking** of the assignments is compiled.
- Comproved's underlying **algorithm** selects a **new pair** to be compared each time, ensuring that every 'work' is compared a certain number of times. Depending on the configuration, if you have twenty works and five assessors, they would need to compare thirty pairs each to achieve a comparison rate of fifteen pairings per work.
- Since these comparisons are done holistically and each comparison usually takes **no more than three minutes**, the total time it takes for one assessor to compare twenty papers is about ninety minutes.

### (2) Workload reduction through peer assessment (summative) – UHasselt

Looking for ways to reduce the **marking workload**, a lecturer at UHasselt decided to use Comproved for **peer assessment**. A **hundred students** were asked to write a **paper** and upload it to Comproved. The algorithm generated **random pairs** of papers, which the students had to compare. They also had to give feedback on every paper. This resulted in a **ranking** of the papers, as well as **feedback** from a dozen fellow students per paper.

Prior to this exercise, in the formative assessment stage, the lecturer had already marked the papers with a **pass/fail** system. Papers that satisfied the final competence received a 'pass', the rest got a 'fail'. The lecturer also gave feedback on each paper.



### What makes comparative assessment effective?

- Comparisons are **easy and quick**.
- The expertise of **multiple assessors** is used.
- Each work is **compared multiple times**.
- Scientific research shows that this leads to **reliable and valid assessments**.
- Comparative assessments in Comproved can be organised per study programme or institution-wide, and even across **institutions and programmes**.

### When can comparative assessment be used?

- **(Peer) assessment** (formative and/or summative): mood boards, academic writing, internships, (e-)portfolios, self-reflection, argumentative texts, mathematical problem-solving, interactive installations (live assessments), etc.
- **Selection**: CV screening, project proposals, etc. Ultimately, this is about choosing a 'winner'.
- **Professionalisation**: assessor training for Examinations Board members (writing skills). The results from Comproved can be used to professionalise assessors and to streamline and harmonise their assessment practices.

### What are the advantages of Comproved in peer assessment and peer feedback?

- Using Comproved, students assess the work of others comparatively, thereby learning to give **more targeted feedback**.
- Students learn from seeing tasks of **varying quality** submitted by their peers. The many examples they see provide plenty of inspiration. Moreover, seeing the work of others **stimulates reflection on one's own work**. Having to give specific feedback to peers helps **to master quality criteria**.
- Both **learning from examples** and **comparing** and looking for similarities and contrasts are proven learning principles (e.g. Carless & Cham, 2016; Pachur & Olsson, 2012).

- Even though fellow students aren't experts (yet), they're often good judges of whether someone else has done a good job or not. Comparing two works and identifying the best one is easy, so **no extensive instruction** or training is needed beforehand.
- Moreover, students consider the **feedback** they get from their peers **rich and valuable**. They accept the results because they have faith in one another's competences. And they're **prepared to act on the feedback** received, precisely because they just gave feedback themselves (expecting that this will encourage their peers to reflect and take action).
- Both the **general discussion** of the assignment afterwards and the **specific feedback** given are **valuable** and help to clarify the criteria for both lecturers and students. The great thing about this is that **students get a say** in the **assessment criteria**. This ensures their understanding and acceptance of these criteria.
- Comproved provides lecturers with a lot of **extra information**, as they can monitor to what extent students have similar views on what constitutes high quality. Lecturers can also monitor how long students took to complete the assessment, which helps to analyse the reliability of the assessment.

### What should you bear in mind when using Comproved?

- It's less suitable for the assessment of **lengthy papers** (dozens of pages), because then the workload of comparing everything multiple times is simply too high.
- You need at least two assessors, but preferably **at least four**.
- You need to **configure the settings beforehand**: the type of assignment, comparison questions (e.g. 'Which one is better?'), the number of comparisons, whether or not to include feedback, etc.



- It's best to use the tool in a **formative** assessment first, before using it for **summative** assessments.

**What can be learned from the experiences of Comproved users?**

- Students experience the platform as a '**safe environment**', as the assessments are completely anonymous. Nobody knows who wrote which paper.



## Want to know more?

### Criteria and rubrics

Centre of Expertise for Higher Education (2013). *Vijftig onderwijstips*. Antwerpen-Apeldoorn: Garant. (Available online to UAntwerp staff [here](#)) > Teaching Tip 'Rubrieken als begeleidings-en beoordelingsinstrument' (info in Dutch)

Bloxham, S. (2009). Marking and moderation in the UK: false assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, 34(2), 209-220. <https://doi.org/10.1080/02602930801955978>

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159-179. <https://doi.org/10.1080/02602930801956059>

### Pairwise comparison

Coertjens, L. Lesterhuis, M., Verhavert, S., Gasse, R., & De Maeyer, S. (2018). *Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering*. Retrieved on 21 May 2021 from <https://www.kennisdelingtaalbeleid.org/wp-content/uploads/2018/01/Teksten-beoordelen-met-criterialijsten-of-via-paarsgewijze-vergelijking.-Een-afweging-van-betrouwbaarheid-en-tijdsinvestering.pdf> (info in Dutch)

Mortier, A. V., Lesterhuis, M., Vlerick, P., & Maeyer, S. D. (2015). Comparative judgment within online assessment: exploring students' feedback reactions. In E. Ras & D. J. Brinke (Eds.), *Computer Assisted Assessment. Research into E-Assessment* (pp. 69-79). Springer International Publishing.

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2), 157-170. <https://link.springer.com/content/pdf/10.1007/s10798-011-9189-x.pdf>

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281-300. <https://doi.org/10.1080/0969594X.2012.665354>

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273-286. <https://doi.org/10.1037/h0070288>

### Comproved

Comproved. (no date). *FAQ*. Comproved: Assess better. Learn more. <https://comproved.com/en/faq>

i-Learn Vlaanderen. (2021, 18 January). *Onder de motorkap: Comproved*. i-Learn Vlaanderen. <https://www.i-learn.vlaanderen/nieuws/edtech/onder-de-motorkap-comproved> (info in Dutch)

Deneire, A. (2017, 10 September). *Zo laat je je studenten lessen trekken uit krachtige D-PAC-feedback*. Edubron blogt. <http://www.edubronblogt.be/onderzoek/zo-laet-studenten-lessen-trekken-krachtige-d-pac-feedback> (info in Dutch)

Centre of Expertise for  
Higher Education (ECHO)  
Venusstraat 35  
B - 2000 Antwerp  
[echo@uantwerpen.be](mailto:echo@uantwerpen.be)  
[www.uantwerpen.be/echo](http://www.uantwerpen.be/echo)





Willems, T. (2020, 30 November). *Laat je studenten elkaars werk rangschikken met Comproved*. Avans Hogeschool. <https://tools.avans.nl/articles/laat-je-studenten-elkaars-werk-rangschikken-met-comproved> (info in Dutch)

### User experiences with Comproved, via (social) media and/or contact with Comproved

Twitter. (no date). *Comproved*. <https://twitter.com/comproved>

Facebook. (no date). *Comparative judgement to the rescue*. <https://www.facebook.com/groups/2393764744190491>

De Wilde, B. (2019, 7 September). *Hoe evalueer je competenties?* Klasse. <https://www.klasse.be/192665/hoe-evalueer-je-competenties/> (info in Dutch)

### Centre of Expertise for Higher Education (ECHO): Teaching Tips

[Creativiteit stimuleren en beoordelen](#) (November 2015)

[Measure what you want to know](#) (November 2013)

[Peer assessment](#) (September 2013)

[Reliability of peer assessment](#) (June 2017)

[Rubrieken als begeleidings- en beoordelingsinstrument](#) (September 2017)

### General

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2011). *Building a Validity Argument for the Test of English as a Foreign Language*. New York: Routledge.

Laming, D. (2003). *Human judgment: The eye of the beholder*. Andover: Cengage Learning EMEA.

McNamara, T. F. (1996). *Measuring second language performance*. Addison Wesley Longman.

Carless, D., & Kam Ho Cham, K. (2016). Managing dialogic use of exemplars. *Assessment & Evaluation in Higher Education*, 42(6), 930-941.

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, 65(2), 1-34.

### UAntwerp-specific inspiration sources

UAntwerp has a university-wide Comproved licence. On [Blackboard](#), you can find specific (technical) information to get started with Comproved. A user guide for students is also available (info in Dutch).

The [Education Info Centre](#) has compiled some good practices for peer assessment, including for pairwise comparison (info in Dutch).

The 'On-campus and online education' page on Pintra contains a lot of information on [peer assessment and peer feedback](#) (info in Dutch).