Trialling a Comparative Assessment Tool for Modelling-oriented Assignments – Findings and Recommendations

Research-in-Progress

Andreas Drechsler
School of Information Management
Victoria University of Wellington
Wellington, New Zealand
Email: andreas.drechsler@vuw.ac.nz

Abstract

This paper reports on the findings of trialling an assessment tool (D-PAC) which implements a comparative marking algorithm and promises to be superior to traditional rubric-based marking. The tool was trialled with 18 student submissions of an enterprise architecture modelling assignment from an undergraduate course, and the trial results were compared with the regular rubric-based marking for the same assignment. The tool was found to be easy to use and the assessment outcome yields some interesting differences. Based on these experiences, the paper derives recommendations for future uses of the tool for modelling-oriented assignments in Information Systems or Computer Science courses. Course coordinators and other decision-makers in universities can draw on these recommendations make an informed choice of whether to consider changing their marking approach for their courses or to introduce the tool as part of the applications they provide for their academics.

Keywords computer-based assessment, modelling-oriented assignments, comparative judgment, d-pac

1 INTRODUCTION

Student assessments are an integral part of university courses. Among other things, they serve the purpose of evaluating how well the students acquired the intended knowledge or competences as set out in the course learning objectives (Biggs 1996). Marking these student assessments is therefore an integral part of running a course. During this process, marking criteria, schemes, or rubrics are commonly used to convey the lecturers' expectations to the students and, subsequently, to assess the assignment work the students submit (Jonsson and Svingby 2007). However, rubric-based marking has certain limitations to giving a sound holistic judgment of student performance (Sadler 2009).

An alternative approach to rubric-based marking is comparative marking (Pollitt 2012), based on the Law of Comparative Judgment (Thurstone 1927). In a nutshell, instead of using criteria-based evaluation, pairs of assignment submissions are compared and judged 'which one is better?' by a number of assessors until a reliable ranking from best to worst submission is achieved. Comparative marking promises to provide superior results compared to rubric-based marking and has been implemented in a web-based software tool (Coenen et al. 2018; D-PAC 2019).

Assessments in Information Systems (IS) courses can take a range of shapes and formats such as essays (e.g. for case-based assignments), working code / prototypes (e.g. for software development assignments), or graphical models and diagrams (e.g. for business process or enterprise modelling assignments). Moreover, these assignments can have a rather limited and focused scope (e.g. centered on the mastery of a single or a few concepts or programming / modelling language constructs), or be of a wider and more open nature (e.g. a coherent case analysis essay, a complete software prototype including documentation, or a comprehensive enterprise model report including model descriptions). The former assessment types are often of a more formative nature, whereas the latter often fulfil more summative assessment purposes. Both, however, can be – and often are – assessed using rubrics.

With this range of possible assessment types, the question arises which assessment types could benefit from comparative marking instead of rubric-based marking. Relevant criteria to answer this question are the quality of the marking outcomes (marks / grades as well as feedback for the students) and also the necessary effort to achieve these outcomes, as there is usually a trade-off between these two criteria. Here, Coenen et al. (2018) report a considerably high level of reliability for a wide range of assessment types (including textual essays and entity relationship (ER) modelling) with comparative marking involving multiple assessors, and Coertjens et al. (2017) also report substantial time saving effects when evaluating textual essays. However, to the author's knowledge, the assessment of more complex modelling assignments beyond ER diagrams is not covered in the literature on comparative marking.

This paper closes this gap and provides an experience report of assessing a rather complex student assignment – a comprehensive enterprise architecture (EA) report produced in a 3rd year undergraduate EA course – in a trial with the aforementioned comparative marking tool called D-PAC. The trial results allow to assess D-PAC's viability as a marking tool and also to give recommendations on the possibilities and limitations regarding its further use in IS and other courses.

The paper is organized as follows. The second section discusses rubric-based marking and comparative marking and introduces the D-PAC tool. The third section covers the evaluation context and methodology employed in the D-PAC trial. The fourth section presents the evaluation's findings, and the fifth section derives recommendations, along with a conclusion and an outlook.

2 BACKGROUND

This section discusses the rationale behind rubric-based versus comparative marking approaches, gives a brief overview about its underlying Law of Comparative Judgment, and introduces the D-PAC tool which implements a comparative marking algorithm.

2.1 Rubric-based Marking: Benefits and Drawbacks

Using marking rubrics to explicate the criteria that are used to evaluate student performance for a particular assignment is a common practice in universities. Rubrics make the lecturer's expectations explicit, thus provide guidance and an opportunity for self-assessment for students during their work on an assignment. Rubrics also allow a more reliable scoring compared to no rubrics if multiple assessors are involved (Jonsson and Svingby 2007). However, rubric-based marking also has a number of limitations (Sadler 2009): Due to the focus on the specified criteria, rubrics may emphasise an evaluation of certain parts of an assignment over a more holistic assessment. Along similar lines, an assignment submission holistically perceived as 'brilliant' may not even score highly on any criterion. The criteria may also not be sufficiently distinct from each other. And since it is not practical to have a

large number of criteria in a rubric, the chosen criteria often only paint a partial picture of the knowledge or competences that are to be assessed. Lastly, different assessors may interpret the same criteria differently, leading to reliability issues when more than one assessor applies a rubric to an assessment.

2.2 Comparative Marking versus Rubric-based Marking

An alternative approach to rubric-based marking is comparative marking (Pollitt 2012). Comparative marking has its roots in the Law of Comparative Judgment (Thurstone 1927). This law postulates that a person can quickly assign a 'value' to a perceived phenomenon (such as an assignment submission), and when asked to compare two similar phenomena (= submissions), their 'values' get compared quickly, thus enabling a quick assessment of 'which is better?'.

Building upon this law, comparative marking algorithms identify assignments of similar quality for comparison and therefore reduce the number of necessary comparisons while retaining a certain reliability (Pollitt 2012). In other words, there is a trade-off between quality (reliability) and time (efficiency). Coenen et al. (2018) implemented such an algorithm in a web-based tool (D-PAC 2019) and evaluated it in eleven settings in different contexts and with different assessment types. They found that their implementation promises to be a reliable implementation of comparative marking that simultaneously provides superior assessment results compared to rubric-based marking. They also provide an indication that marking textual essays with the tool is more efficient than using rubrics.

2.3 D-PAC: The Tool Used for Comparative Marking

D-PAC (2019) is a web-based assessment tool that implements a comparative marking algorithm.

To set-up an assignment for evaluation in the tool, the number of student submissions (N), the number of assessors (A), and the intended reliability (.6 to .7 or .7 to .8) needs to be specified. The tool vendor recommends a minimum of 3 assessors. A higher desired reliability requires a higher number of comparisons per assessor, namely (N * 7.5)/A for the former, and (N * 10)/A for the latter reliability threshold. After the tool is set-up, assessors can log into the tool and start comparing the student submissions (Figure 1). After each comparison, there is a second screen with two text boxes per submission for feedback to the students – one for strengths and one for areas for improvement.

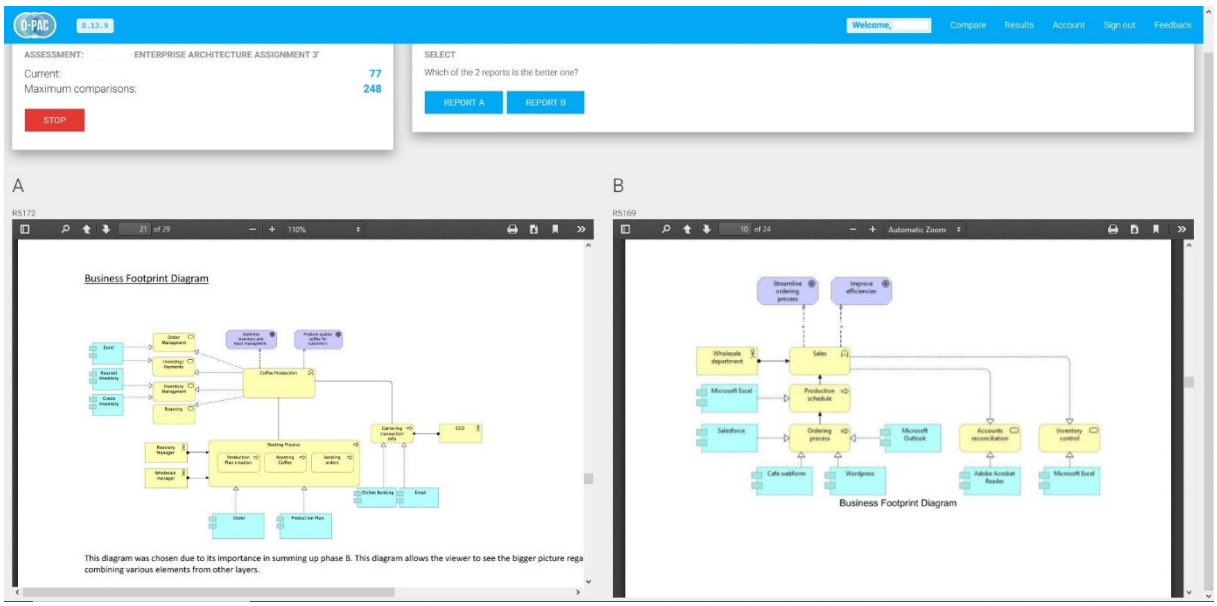

Figure 1. Comparison screen

By clicking on 'Results' at the top, an assessor can start view the current ranking of submissions from best to worst based on the assessors' assessments as calculated by the underlying algorithm (Figure 2). The x-axis contains the assignment submissions, and the y-axis the relative position based on the submission quality. The corresponding numerical value is called 'ability' in the tool exports. The vertical lines indicate the confidence intervals for the ability score. Clicking on a blue dot shows the corresponding submission and the feedback given by each assessor.

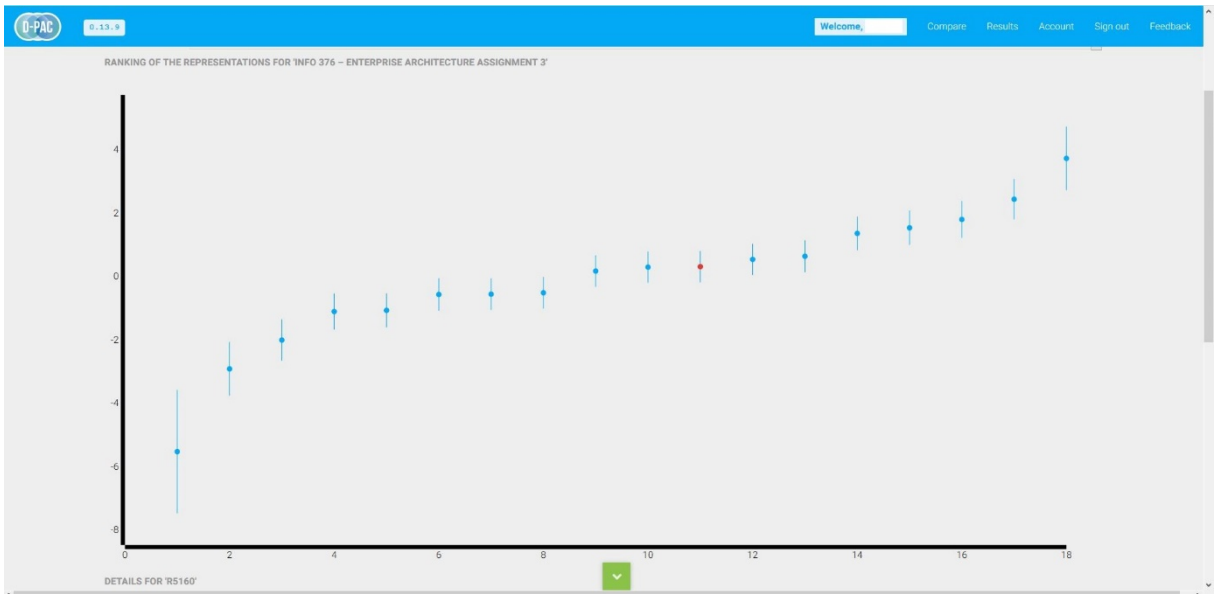

Figure 2. Ranking screen

3 EVALUATION CONTEXT AND METHODOLOGY

This section now covers the context and methodology by which the D-PAC tool was evaluated for the use in IS modelling-oriented assignments.

The evaluation took place in a 3rd year undergraduate Enterprise Architecture (EA) course. The assignment that was chosen for the evaluation was a comprehensive baseline (or as-is) EA report for a part of a fictional company. The report comprised the main TOGAF layers (vision, business architecture, information systems architecture, technical architecture) (The Open Group 2019a). For each layer, the students had to create a number of suitable ArchiMate models (The Open Group 2019b), other architectural TOGAF artefacts (such as catalogues and matrices), and accompanying descriptions for each artefact to produce a coherent report. The number and choice of architectural artefacts was up to the students, who worked in teams of three or four.

The assignment had summative and formative aspects to it. The report was the first coherent EA report the students had to produce in the course beyond a focus on the individual architectural artefacts; it therefore evaluated the students' EA modelling proficiency in a summative way. Simultaneously, this baseline EA report was the foundation for a subsequent target EA report that was about covering a proposed change to the EA to address an organisational problem. Therefore, the report also doubled as a formative assignment for the competence to produce coherent EA reports.

There were 18 student submissions to be evaluated by D-PAC. Regular rubric-based marking of these submissions had taken place during the EA course, and only the results of this rubric-based marking were used to determine the students' performance in the course. The D-PAC evaluation was run separately. The course coordinator and author conducted the regular rubric-based marking and, with a two months' time difference after the conclusion of the trimester, was one of the two D-PAC assessors. The other assessor was a student teaching assistant who had taken the EA course in the year before. This teaching assistant was not involved in the regular marking of any EA course assignments but had some marking experience in other courses with modelling content.

Since the evaluation was a trial, the vendor set up the assessors in the tool and imported the student assignment submissions manually. The assessors kept track of the time they spent, and also provided qualitative feedback about their marking experiences. In the end, the final ranking of the student submissions in D-PAC was compared to the order of the student submissions in the university's learning management system created by the scores achieved by regular rubric-based marking during the course.

Note that, while the vendor recommends three assessors as a minimum for a valid and reliable outcome (see section 2.3), there was an issue with how the vendor set up the trial, which lead to the teaching assistant assessor doing considerably more comparisons than required (and budgeted for). After the author finished his assessment and noted this issue, the vendor assured that an "excellent reliability" of .87 has been achieved anyway, so that there was no need for a third assessor in this regard.

4 FINDINGS

This section presents and discusses the quantitative and qualitative findings of the tool evaluation.

4.1 Quantitative Findings

The first quantitative finding concerns the **time** used for marking. With D-PAC, marking essentially takes place in two phases. In the first phase, each comparison involves two submissions that have not been compared previously. Hence, in this phase, the comparison takes considerable time for two reasons. First, it is the first time the assessor encounters each student submission and therefore needs to read and evaluate them carefully, in order to make a well-informed decision on which one is better than the other. Second, it is also the first time the assessor has the opportunity to give written feedback on both submissions. After each submission has been compared at least once, the second phase begins where two previously compared submissions are compared with one another, based on the underlying algorithm. In this phase, an assessor becomes increasingly familiar with the submissions and the focus can shift towards judging the superiority of one submission over the other. Previously given feedback is shown after an assessment was made and can be edited or amended.

Table 1 shows the difference in marking times per submission for comparative versus rubric-based marking for the experienced and the less experienced assessor. Compared to rubric-based marking, the first phase takes less time since the feedback is guided only by the assessors' holistic impression instead of rubric-based criteria and therefore tends to be less detailed (see also next section). Times varied much less compared to rubric-based marking where more issues in a student submission usually warrant a more in-depth assessment along with more in-depth feedback for each rubric. Note that there is no comparison time for rubric-based marking and a less experienced assessor in Table 1 since there were no comparable instances in the regular course where a less experienced assessor conducted rubric-based marking including writing feedback.

	Comparative marking (1st phase)	**Comparative marking (2nd phase)**	**Rubric-based marking**
Experienced assessor	~8 min incl. feedback	~1-2 min	~10-25 min incl. feedback
Less experienced assessor	~15 min incl. feedback	~5-7 min	n/a

Table 1. Comparison of marking time per submission

The second quantitative finding concerns the **number of assessors** and the **number of assessments per assessor**. At the author's school, it is not uncommon to teach a course with an academic colleague and having one or several student tutors for undergraduate courses who often also have some marking responsibility as part of their duties. Therefore, having three or four assessors for large undergraduate courses would be feasible. This may not be the case, however, for smaller graduate courses or different teaching arrangements. There, using comparative marking would require additional resourcing. For the number of assessments per assessor, comparative marking scales reasonably well. While the formulas for the numbers of comparisons given in section 2.3 contain a multiplier of 7.5 or 10 for each additional submission, only the first phase for each assessor (until all submissions have been compared and feedbacked at least once) is truly time-intensive (see also Table 1). The comparisons afterwards take much less time in comparison, and therefore lessen the impact of the multiplier.

The **comparison of the ranking** achieved by comparative marking with the ranking produced by the numerical marks scored by rubric-based marking also yield interesting insights. 10 out of 18 submissions achieved a similar rank (either the same rank or with up to plus or minus 2 difference), 4 submissions placed considerably better (3, 4, 4, and 5 places) and 4 placed worse (3, 4, 6, and 12 places) compared to rubric-based marking. All four worse rankings can be traced to criteria that the marking rubrics were ambiguous on or did not contain at all: First, for the 12 place outlier, one assessor valued the existence of EA catalogues and matrices considerably more highly than the other – and the submission, with otherwise great diagrams and writing, had none. Second, the holistic assessment of comparative marking caught a decrease of writing quality and quantity towards the end of some of the other worse ranked reports. Third, most of the worse ranked submissions had noticeably plain or 'bad' layouts and diagram quality. This particular reason (especially the submissions' layout) also impacted the higher rankings compared to rubric-based marking to an extent. Here, both assessors reported being influenced positively by 'nice' fonts, page layouts, and colour schemes. The second reason for a higher ranking is a reduced attention to issues regarding the EA modelling and diagram details in the comparative marking approach. In other words, the formative aspect of the assignment – the ability to

write coherent and appealing EA reports – was over-emphasised in comparative marking, whereas the summative aspect of the assignment – the ability to produce 'correct' EA catalogues, matrices, and diagrams – was better taken into account while following rubrics.

4.2 Qualitative Findings

Regarding the **marking process**, the less experienced assessor found it somewhat difficult to adjust to this type of marking, based on her previous marking experience. Both assessors found themselves liking those submissions which had a format that was easier to read and follow, which made both more inclined to choose one submission with a more attractive format/layout over another. Both assessors therefore felt to be more susceptible to having a bias, especially after encountering the same reports over and over. The more experienced marker tended to use 'scope' (breadth / depth of architectural artefacts on each layer) as the emergent main guiding principle to quickly decide on better/worse. This is a shallower perspective compared to rubric-based marking and would emphasise the quantity of architectural artefacts over their quality or necessity. Simultaneously, both assessors found this type of marking to be a lot more flexible as they could base their decision on factors that are outside of any rubric. However, while this marking was simpler, both assessors found themselves also second guessing some decisions after giving feedback or reading the previously given feedback after a comparison.

The **type of feedback** that the assessors gave also turned out to be different. The more holistic assessment perspective means less detail-oriented feedback since there is no further guidance – such as rubrics – for the assessor. In other words, a comparative perspective works well from a 'report customer' point of view, but less so from a lecturer's perspective that wants to ensure that also the 'little errors' (e.g., adherence to modelling rules or guidelines) are identified and given feedback on. The option to add or change feedback after each comparison compensates a little bit for this issue, however, the emphasis on quick comparisons especially in the second marking phase works against an increased attention to detail by the assessors. This is further exacerbated by feedback almost becoming an instrument of confirmation for a choice at some point. In contrast, carefully going through every rubric element for each assignment in rubric-based marking can lead to very specific and focused feedback.

Lastly, both assessors found the **D-PAC tool** to be very straightforward and easy to use. The side-by-side presentation of two submissions (Figure 1), however, sometimes led to legibility issues if there are smaller fonts in the report (text, figures, diagrams). While the files can be opened in a PDF reader application from the PDF viewer embedded in the browser, doing so removes the ability to do an easy side-by-side comparison. It is also not possible to change your decision after making the choice which assignment you thought was better, and there were times for both assessors where they wanted to change their initial decision, particularly after writing feedback. While the order of comparison first, feedback second makes very much sense, it would be nice to have a 'notepad' in the tool to write down considerations during the comparison period that would not be shared with the students, but visible during the feedback phase. Finally, there was very limited documentation and guidance for the decision-maker, configurer, and first-time user at the time of the set-up, due to the tool being still in development. The vendor, however, was very responsive regarding inquiries, the trial set-up, the issue with the wrong number of target comparisons per assessor, and providing raw exports for a deeper analysis.

5 RECOMMENDATIONS, CONCLUSION AND OUTLOOK

This section draws on the previously reported findings to give more general recommendations regarding the potential use of D-PAC or similar comparative marking tools to assess student assignment work.

The first thing should be to have **clear goals** in mind for choosing a comparative marking approach for an assignment. As discussed above, the comparative approach is more suited for holistic assessments of assignments that are open to a degree and require some form of creativity from the students. This is regardless whether the assignments are formative or summative in nature. Considering **the course and assignment context**, an absence of familiar rubrics for a somewhat open assignment may also increase the students' perceived ambiguity even further, especially when they are used to rubrics and less open assignment styles. The lack of a rubric would also require some other means of conveying the expectations both to the students and to the assessor team. According to the vendor, future versions of the tool will offer configurable criteria for the feedback form. A key issue both assessors reported during the marking was the danger of being unduly influenced by clean layouts (over 'substance'). A solution could be mandatory document templates provided for the assignment – but those would work against an intended openness by creating a certain uniformity among the submissions. Overall, choosing a comparative marking approach requires a **close alignment** with the intended learning and assignment objectives, the course context, the assignment type and design, and the student expectations.

Another aspect to consider is adequate **resourcing** for the marking process. While comparative marking may be reasonably time efficient, the vendor guidelines strongly recommend three or even four assessors, which may not be available in all cases. Comparative marking can therefore well be more resource intensive and less flexible in terms of resourcing than rubric-based marking. An advantage is that all assessors can work and give feedback simultaneously in the tool, without having to worry about coordinating marking times or assignments. Also, the D-PAC tool is a standalone tool at the moment, and the student assignments and assessors have to be set-up manually by the vendor.

Lastly, to conform with established expectations and protocols how course assignments need to be assessed, comparative marking may require an **additional step** of assigning a numerical mark or letter grade to ranked reports after finishing the work in the tool. While the tool provides an 'ability' score, it does not map easily to a scale of 0-100 (where 50 commonly is the passing threshold) or letter grades A-F. This step of essentially 'making up' boundaries (e.g., what does a B+ look like?) therefore requires special consideration by the assessors. Such a step would also allow for a **'moderation' check** of the final ranking to catch assignments that ended up being ranked too highly or lowly due to differing assessor perceptions (e.g., the 12 place 'outlier' as discussed in section 4.1). Compared to rubric-based marking, the need for post-marking moderation therefore does not go away, it just takes a different form.

Of course, this trial was not without **limitations**. First, it was limited to a single assignment in a single course in a single year. More in-depth assessments are needed of how comparative marking fares over time, for different assignment types, and different topics. There was also a lack of integration into a course – both upfront regarding managing student expectations without a rubric, and also after the marking process to come up with a value that integrates with established grading protocols. Finally, the issue of having a erroneously too high target number of comparisons set up effectively limited the trial to two (instead of three) assessors. Luckily, the high reliability between the two assessors prevented this issue from having more severe impacts on the validity of the trial. A third assessor could have potentially provided more and different insights, however.

In sum, this trial of a tool for comparative marking found that comparative marking does enable the assessors to provide insight into aspects that are not necessarily covered in a rubric and to take a holistic 'customer' perspective. There is also the potential to realise time efficiency gains without sacrificing assessment validity due to not having to go through each rubric for each submission. In contrast, rubric-based marking allows a more detail-oriented evaluation of an assignment and therefore may align more easily with the criteria that needed to be achieved as per given course learning objectives or an assignment brief. The key conclusion therefore is that, for certain assignment types, comparative marking and the D-PAC tool show considerable promise and warrant further attention and trials.

6 REFERENCES

Biggs, J. 1996. "Enhancing Teaching through Constructive Alignment," *Higher Education* (32:3), p. 347.

Coenen, T., Coertjens, L., Vlerick, P., Lesterhuis, M., Mortier, A. V., Donche, V., Ballon, P., and Maeyer, S. D. 2018. "An Information System Design Theory for the Comparative Judgement of Competences," *European Journal of Information Systems* (27:2).

Coertjens, L., Lesterhuis, M., and Verhavert, S. 2017. "Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering," *PEDAGOGISCHE STUDIEN* (94:4), pp. 283–303.

D-PAC. 2019. "D-PAC - English," *D-PAC*. (https://www.d-pac.be/english/, accessed July 29, 2019).

Jonsson, A., and Svingby, G. 2007. "The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences," *Educational Research Review* (2:2), pp. 130–144.

Pollitt, A. 2012. "Comparative Judgement for Assessment," *International Journal of Technology and Design Education* (22:2), pp. 157–170.

Sadler, D. R. 2009. "Indeterminacy in the Use of Preset Criteria for Assessment and Grading," *Assessment & Evaluation in Higher Education* (34:2), pp. 159–179.

The Open Group. 2019a. "The TOGAF® Standard, Version 9.2." (http://pubs.opengroup.org/architecture/togaf9-doc/arch/, accessed July 29, 2019).

The Open Group. 2019b. "ArchiMate® 3.0.1 Specification." (http://pubs.opengroup.org/architecture/archimate3-doc/toc.html, accessed July 29, 2019).

Thurstone, L. L. 1927. "A Law of Comparative Judgment.," *Psychological Review* (34:4), p. 273.