Chapter 8
Oranges and Apples? Using Comparative Judgement for Reliable Briefing Paper Assessment in Simulation Games

Pierpaolo Settembri, Roos Van Gasse, Liesje Coertjens, and Sven De Maeyer

8.1 Introduction

One of the aspects of simulations that have traditionally been neglected by the relevant literature is their supposed added value as a teaching tool.[1] This chapter deals with assessment as well, yet its focus is not on simulations as such but on those who take part in them, i.e. the participants. This is clearly not an entirely separate debate because, ultimately, it is on the impact that the simulations have on their participants that one has to judge the pedagogical value of simulations themselves. Only if simulations are designed with a rigorous assessment framework which, in turn, is

Pierpaolo Settembri writes in a personal capacity and the views he expresses in this publication may not be in any circumstances regarded as stating an official position of the European Commission.

[1] Raymond and Usherwood (2013, p. 4) put it extremely clearly: "University faculty must ask themselves what a simulation adds to a student's knowledge base that cannot be learned more efficiently in a traditional classroom setting, and how this can be measured". Baranowski and Weir (2015) offer a deep review of the literature evaluating the effects of simulations, coming to the conclusion that "a small but growing body of evidence lends support to the contention that students who participate in simulations do in fact learn more than students not taking part in this exercise". For a different outcome, see Raymond (2010).

P. Settembri (✉)
European Commission, Brussels, Belgium
e-mail: pierpaolo.settembri@ec.europa.eu

R. Van Gasse · S. De Maeyer
University of Antwerp, Antwerp, Belgium
e-mail: roos.vangasse@uantwerpen.be; sven.demaeyer@uantwerpen.be

L. Coertjens
University of Antwerp, Antwerp, Belgium

Université Catholique de Louvain, Louvain-la-Neuve, Belgium
e-mail: Liesje.coertjens@uclouvain.be

P. Bursens et al. (eds.), *Simulations of Decision-Making as Active Learning Tools*, Professional and Practice-based Learning 22,
https://doi.org/10.1007/978-3-319-74147-5_8

adequately reflective of predefined learning objectives can they bring tangible and measurable benefits. Rather than engaging in an abstract discussion as to whether simulations are better or worse teaching tools than other more traditional methods, this chapter offers some insights on how to ensure that participants in simulations are assessed fairly and thoroughly. One key suggestion is to rely on comparative judgement to assess written assignments that are produced in the context of a simulation, particularly when the number of participants is elevated.

It is not unusual, indeed, that the assessment of simulations is (also) based on the materials that participants are required to prepare as part of the simulation activity, in addition to the evaluation of their performance in the exercise. Participants are often given tasks to perform ranging from writing (short) papers to keeping a diary on the experience. In this respect, Chin et al. (2009) even suggest that the teacher should involve a second person who can give advice and feedback on the tasks carried out during the activity without excessively distorting the behaviour of the participants.[2] We emphasise this type of assessment based also on the materials prepared in the context of a simulation because this is the method that was used in the exercise described in this chapter.

In the next sections, we will attempt to address some of the recurrent difficulties related to evaluating participants in simulation games on an individual basis. We will do so by, first of all, looking at those specificities of simulations and challenges that make them less amenable to individual grading. Secondly, based on a concrete example, we will describe the ways in which these difficulties have been tackled. In this context, we will pay special attention to comparative judgement, a method that compares performances two by two, instead of assessing them one by one, which turned out to be particularly suitable for our purposes. In the conclusion, we will consider the broader implications of these insights for the assessment of simulations.

8.2 The Challenge

Providing an individual grade to participants in simulation games is one of the most difficult tasks simulation instructors face. Such is the challenge that it is not uncommon to find simulation games where participants can only pass or fail.[3]

At a basic level, the issue is simple: the simulation game is a collective exercise, and assessing the performance of its individual participants is per se counterintuitive. Yet, individual grading is essential not only for practical pedagogical reasons but also as an incentive for participants to play the simulation realistically: real negotiators do have an individual interest to perform well, on top of a collective interest to secure an overall acceptable result. The prospect of an individual assess-

[2] Alternatively, the data for the assessment can be collected by videotaping the meetings. Although this is a highly intrusive method, it yields material useful for subsequent analyses.

[3] Perchoc (2016) mentions the example of the International Relations Department of the College of Europe in Bruges.

ment is a strong motivating factor that is likely to enhance the commitment and hence the performance of participants. Conversely, an exercise in which students are assessed only for their collective output entails the risk that some participants will take a more passive stance and benefit from the motivation and activism of their more proactive peers.

But why is it so difficult to grade participants?

In our view, there are two main reasons:

1. The first reason has to do with the roles participants are assigned in the context of the simulation, which are not necessarily equivalent. Because in most negotiations some players have more important roles than others, participants performing these roles in a simulation are clearly advantaged by the greater responsibilities, resources, opportunities and exposure they have. Conversely, participants with minor roles have less stakes, ammunitions and occasions to shine.
2. Secondly, like in real negotiations, there is more than meets the eye also in simulation games. What determines the outcome of a negotiation is the result of formal and informal dynamics, of visible and invisible activities, of intentional actions and unintended consequences. Attributing the credit or the responsibility to individual participants for the success or the failure of a negotiation is thus a very risky task[4] and the same applies to simulations. Moreover, the challenge clearly increases with the number of participants, where it becomes even more difficult to keep track of key developments.

8.3 Some Solutions

How can these challenges be addressed? Prior to any other considerations, the instructor needs to clearly define the learning objectives against which the performance of participants should be assessed.

There is some variation among scholars as regards the skills simulations are expected to impart to participants. This is not surprising or problematic: like traditional courses, simulation games can be versatile tools, which can be used to facilitate the acquisition of different skills. What is important is that these skills – or learning objectives to be attained – are defined beforehand and are accompanied by an assessment framework that is adequate to ascertain whether and to what extent they have been acquired.

For Raymond and Usherwood (2013), simulations typically aim to achieve one or more of the following learning outcomes: substantive knowledge acquisition, skill development (e.g. negotiation skills) or group socialisation. Raiser et al. (2015), on the other hand, classify simulations according to the skills that the organisers

[4] To be noted here that success or failure of a simulation does not necessarily mean that participants managed or failed to find an agreement. This is a subjective notion that the instructor defines on the basis of prior criteria and learning objectives.

want to impart, focusing in particular on the "soft skills" the labour market increasingly expects from graduates. In this respect they identify simulation games focused on *interaction and communication*, others with a focus on *systemic competence* to improve students' ability to deal with complexity, and finally simulation games with a focus on *decision-making and action-related competence*, which put students in situations that train their ability to make decisions, particularly under time pressure, stress and high media attention.

Of course, these objectives are not mutually exclusive, and in fact, most simulations encompass a mix of some or even all of them. This is not without consequences: the more numerous the goals, the more complex the simulation and, in turn, the more challenging the assessment framework – a point that also Raymond and Usherwood (2013) had made clear.

Yet, the reality of most EU simulation games is that, due to their complexity, the two challenges described in the previous section (i.e. role bias and invisible activities) continue to threaten the fairness and the reliability of the evaluation, no matter how well defined are the learning objectives and how zealous and capable is the instructor in assessing them.

To address the first, easier challenge, it could prove helpful to give participants different roles in the course of the simulation, thereby diluting the bias associated with most prominent roles. The logic here is that, for example, the exercise could be divided into two parts, and no participant should have a prominent role in both of them. This is in addition to the standard practice to prohibit students representing their own country, as this could result in an unfair advantage (see, e.g. Obendorf and Randerson 2013[5]).

The second challenge is more difficult to address, but it becomes clearer if we take a step back. As made clear by Raiser et al. (2015), the skills that are tested (and hopefully enhanced) in the context of a simulation game are broader than those that are strictly relevant in formal negotiations, such as the ability to persuade, to build the necessary alliances or to find viable compromises. Equally important are also other skills such as, for example, the ability to understand a problem, identify the relevant information, assess alternative options, devise a strategy and execute it successfully.

The proposed solution is, in a nutshell, a compound assessment framework that combines written assignments and participation. The underlying assumption is that the assessment of individual performance based only on the interactions among participants is difficult and potentially misleading, especially when there are many

[5] In their Model United Nations simulation programme, they have a member of the teaching team to chair the final conference "to maintain equity of opportunity in assessment … and to ensure adherence to the rules of procedure" (Obendorf and Randerson 2013, p. 357). They make a similar exception for the activities of the Secretariat. While there might be an undue advantage granted to those who are assigned these roles (hence the need to mitigate or compensate for it in various ways, as explained in this chapter), similar exceptions could be detrimental to the realism of the simulation itself as it creates an artificial subordination between different categories of players that has no equivalent in reality, as the authors themselves admit.

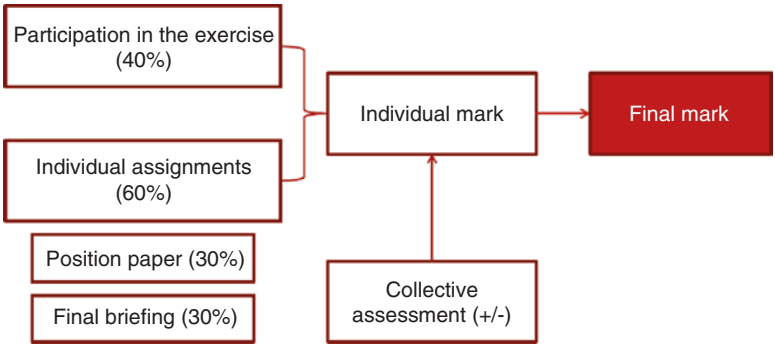

Fig. 8.1 Assessment framework

players. In these cases, individual assignments can be introduced to mitigate this drawback and obtain a more objective basis against which to assess performance.

In a recent simulation exercise offered at the master programme of the College of Europe in Bruges[6] involving around 100 graduate students, the assessment framework proposed to students comprised a combination of individual and collective marks as outlined in Fig. 8.1. This framework reflects the complex mix of skills students attending this obligatory course ran over 1 month are expected to acquire.

It is worth to briefly describe each of these components and to make explicit their link with the skill to be assessed.[7]

The overall "*participation in the exercise*" captures the intensity and quality of the students' contribution primarily, although not exclusively, to the visible part of the negotiation, i.e. the formal meetings (EP committee, COREPER, etc.). It assesses their ability to play correctly yet skilfully with the rules of the game. It reflects the quality of their interaction and the skills each of them displayed in persuading others to move in the desired direction. Although this is still the single most important component of the framework in relative terms, alone it would be not sufficient to determine the final mark.

Among the "individual assignments", the "*position paper*" is a short written contribution (max. two pages) that the students prepare before the start of the negotiation to the benefit of a "client" (a member state, a member of parliament, etc.) that has hired him/her as a consultant to help prepare the negotiating position.[8] It has been introduced to ensure that participants acquire a minimum level of substantive and procedural knowledge about the file under discussion. Given the severe length

[6] The details of this simulation game have been provided in the chapter on verisimilitude. The official page of the course is accessible here: https://www.coleurope.eu/course/settembri-p-hermanin-c-worth-j-negotiation-and-decision-making-eu-simulation-game-50h.

[7] The combination of participation and written contribution is common also to other modules. For example, Obendorf and Randerson (2013) describe a formal assessment based on four components, with a similar articulation: a written country position paper (25%), participation in the simulation (35%), a binder of research sources (25%) and reflective essays (15%).

[8] This assignment has been described in greater detail in the chapter concerning verisimilitude.

limit and the wealth of information available, especially on the Internet, students are obliged to identify, analyse, select and prioritise the information they eventually include in this document.

The "*briefing*", on the contrary, is a more detailed written contribution, based on a specific template that is required in the final stages of the simulation. The briefing is the document that the negotiator receives from the staff to prepare for a negotiation. In the case of this simulation, the briefing is required ahead of, and in preparation for, the final high-level meeting, for example, the European Parliament's plenary meeting expected to vote on a committee report or to approve the outcome of a trilogue with the Council and the Commission. The combination of these three elements, duly weighted, determines the individual mark. Depending on the overall assessment of the simulation (e.g. the realism of its dynamics, the credibility of its outcome, etc.), the teacher could add or subtract the same number of points for all participants, hence determining the final mark.

Compared to the position paper, the briefing is a more elaborate product, which is likely to capture a specific set of skills that the simulation game should in principle strengthen. It is expected that the simulation – through the inevitable interaction, the imperative to prepare and the exposure to peer pressure – will prompt participants to enhance their knowledge and abilities on a number of fronts. It is now time to be more explicit and detailed about these skills:

1. *The topic of the negotiation*. Although this may seem a rather narrow objective – hardly any participant will end up working in the policy area covered by the negotiation – it does yield more general understanding of the complexity that any topic encompasses. It shows, among other things, how supposedly simple issues can become divisive and, ultimately, complex to solve. It also illustrates that, no matter how complex a topic may be, the negotiation will inevitably end up revolving around a limited, but well-defined, set of issues on which positions will polarise. Finally, it will confront the participants with the importance of information, expertise and technical knowledge and the roles that they play.
2. *The dynamics of the negotiation*. The expected gains here are manifold, and any list would not only be incomplete but also subjective. Understanding the relative strength of the players, identifying the most effective channels through which to exert influence, developing the ability to track how a compromise emerged while others failed and seeing persuasion techniques at work are among the most valuable takeaways of a simulation exercise. However, they are not the only ones. The specific nature of EU negotiations will acquaint the participants with the role that the different EU institutions play, and it will show how their institutional position shapes their actual behaviour in the negotiation.
3. *Their own role in the negotiation*. Feeling the impact of your own actions on others (and vice versa) is an extremely powerful learning experience and one that students are unlikely to obtain from other curricula. Each of them will develop their own perception of the simulation and of their individual role in it. In so doing, they will be confronted with the need to (1) identify the problems on the table and the interests that they are defending, (2) prioritise among competing

interests and compromise with other players representing different ones, (3) identify their allies and also the players that they should stay away from and (4) take a certain course of action in pursuit of their objectives. In short, they will *feel* the pleasure of victory and the bitter taste of defeat, and they will experience these emotions personally.

Because of its versatile but also strategic nature, the briefing is expected to capture most of these gains. It should be noted that the briefing is not a post-negotiation report; rather, it is a real tool to support the negotiator, and it is prepared at the peak moment of the negotiation, i.e. before the final round.

The briefing, which is requested with a strict page length limit (max. four or five pages), is structured around five mandatory sections, each reflecting a different skill:

- The "scene setter" (roughly half page), which should reveal the ability to summarise the state of play and the importance of the next meeting/step, from the perspective of the participant
- The "objectives" part (two to three bullet points), which should reveal the ability to prioritise and focus
- The "key messages" section (one to two pages, bullet style), which should reveal understanding of the key issues, the ability to strategise and anticipate the others' moves
- The "defensives" (one to two questions/answers) section, which requires acknowledgement of possible weak points and the development of arguments to counter them
- The "background" (approx. 3/4 of a page), which requires the ability to select and explain key issues under severe space constraints

A compound assessment framework mitigates the challenge described in the previous section, but some difficulties remain, especially as regards the fairness and reliability of the individual mark.

When it comes to assessing "*participation in the exercise*", one hurdle for the instructor is to have a comprehensive overview of what really happened and thanks to whom. Because we know that what happens at formal meetings is not necessarily the full story – but rather the tip of the iceberg – we deliberately encourage participants to share with the instructors all the activities that were not visible during the exercise. Examples may include the report of an informal meeting, the leak of a document to the press and an alternative proposal prepared by a group of like-minded participants. Whereas the first year we ran this simulation game we asked students to notify to the instructors each of these informal (or otherwise invisible) activities, the second year we found it more practical to require that, at the same time as they submitted the briefing, students sent also a one-page activity report setting out all their actions in connection with the negotiation, including on the social media of the simulation.[9]

[9] For a more detailed description of this tool, please refer to the chapter on verisimilitude.

Despite these specific fixes, assigning individual grades is still challenging, particularly when there are many participants. The approach we pursued as regards the assessment of the "participation" has been to identify a reference grade that would reflect the average performance of the group in the course of the simulation and then adjust it upwards and downwards for over- and underachievers, respectively. This required instructors to be attentive at all stages of the negotiation and to take notes at meetings to note down good and less good conducts. In addition, it required an assessment of the activities carried out outside meetings, including on the social media (of the simulation), or shared informally with the instructors because of their confidential/informal nature. This resulted, practically, in an Excel spreadsheet with the names of all participants and, next to each of them, plusses and minuses linked to specific episodes or initiatives.

Assessing the individual *written assignments* of the simulation exercise presented challenges as well. Assessments with a large number of products often require multiple raters. Hence, it is important to assure that students' grades are not impacted by varying severity over time or by differences in rater severity (e.g. scoring the same objects differently at different occasions or dependent on the (good or weak) quality of the prior object that was assessed). Different raters value different aspects of the task, whereby the combination of multiple raters is crucial to achieve sufficient validity in assessments (Bloxham et al. 2016; Pollitt 2012). In particular for more open-ended tasks, it is almost impossible to formulate all relevant criteria in advance (Sadler 2009). Therefore, creating diversity in the view on these tasks by including multiple raters in the assessment process is essential.

Recently, comparative judgement (CJ) has been introduced as an alternative approach to common rating practices in the assessment of complex competences. This method is based on the simple yet crucial assumption that people are better and more reliable in comparing two performances (e.g. briefings) than in assigning a score to a single one (Thurstone 1927; Laming 2003). Figure 8.2 shows the rationale behind the CJ method. Multiple assessors judge a fixed number of random pairs of performances. In each comparison, assessors are asked to select the best performance with regard to the assessed competence (i.e. the one which overall quality is perceived as higher or shows most evidence of ability with regard to the competence).

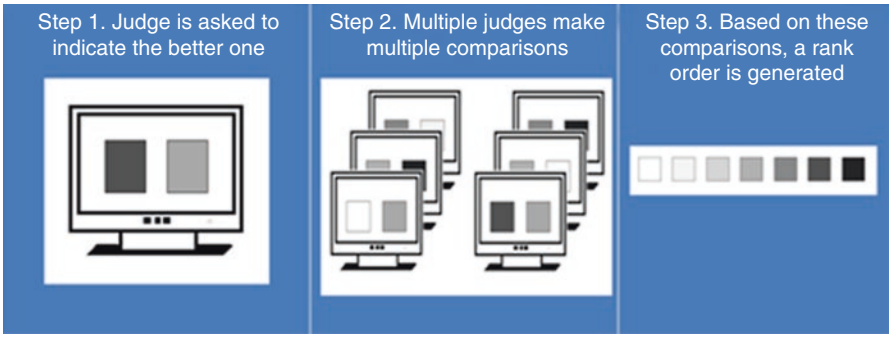

Fig. 8.2 Step-by-step explanation of the CJ method

Note that, due to the fact that assessors only have to choose the better one, the difference in severity between assessors is annulled (Bramley 2007; Whitehouse 2012; Whitehouse and Pollitt 2012). After choosing a performance, the assessor receives a new pair to compare. Several assessors need to make multiple comparisons. Subsequently, by applying a statistical model, all performances are ranked according to the consensus among assessors regarding the quality of performances (Bramley 2007). This rank order is grounded on assessors' intuitive frame of reference in their judgements (Laming 2003).

Up to present, CJ has already been applied to the assessment of a wide range of competences (e.g. mathematics, Jones et al. 2013; geography, Whitehouse and Pollitt 2012; writing, Heldsinger and Humphry 2010, 2013; Pollitt 2012). Each time, using CJ in assessments has resulted in reliable rank orders, with reliability estimates ranging from 0.73 (Jones and Alcock 2014) up to 0.98 (Heldsinger and Humphry 2010). To our knowledge, CJ has not been applied to performances in the political science domain.

8.4 Comparative Judgement Applied to Briefings

In order to investigate the value of the CJ approach for the assessment of simulation products, an assessment was set up using the briefings of 84 students.[10] The assessment was supported by the online *Digital Platform for the Assessment of Competences* (D-PAC) tool. The 84 briefings were judged by four assessors. Previous CJ assessment using the D-PAC tool has indicated that the algorithm used needs 9–15 comparisons per assignment to reach sufficient reliability (i.e. 0.70). Given the low number of assessors, the maximum number of 15 comparisons per briefing was set in the tool. In total, 620 comparisons out of the total possible number of comparisons were completed. The pairs of briefings were drawn randomly from the set of briefing paper having the least completed comparisons at that moment. This approach guarantees that each briefing appears in a pair about the same number of times (e.g. 15 times).

Limitations in the time investment of two out of four assessors resulted in the 620 comparisons being unevenly distributed among assessors. Two out of 4 assessors finished 210 comparisons, and the other 2 assessors finished 100 comparisons. The total duration of the period in which assessors completed their comparisons was 3 weeks.

CJ data is analysed using a Rasch model. Therefore, the scale separation reliability (SSR; Bramley 2015) can be calculated. The measure represents the amount

[10] In fact, the total pool consisted of 96 papers, but 12 of these were of different nature. They were assignments to non-institutional actors (journalists, lobbyists, NGOs and other stakeholders), for which the briefing was not a suitable assignment. These 12 assignments have been assessed separately, but based on the same rationale as in the D-PAC tool. The analysis here focuses exclusively on the larger pool.

of spread in the results that is not due to measurement error (McMahon and Jones 2015). According to Anshel et al. (2013), the SSR provides an indication for how separable the representations are on the final scale of the assessment. Values of the SSR vary between 0 and 1, and a small measurement error (and thus an SSR closer to 1) implies that the relative position of the items on the scale is quite fixed (Andrich 1982). Though research on SSR's characteristics is ongoing, results indicate that it is a good measure for split-half reliability (Verhavert et al. 2016), being the correlation between the two rank orders generated by splitting the assessors randomly into two groups.

For the current study, the SSR was used to evaluate the reliability of the briefing assessment. Furthermore, the evolution of the SSR was examined throughout the assessment (i.e. the increase of reliability at times all briefings were compared 0–15 times). All analyses were conducted using R software.

On average, assessors spent about 7 min to make a decision in a comparison. More detailed analyses of the data revealed however strong outliers for this time estimate (e.g. from 1 h to complete a comparison up to over 3 h, probably due to leaving the assessment open while continuing with other tasks). As such, this average duration is overestimated. This is also evidenced by the fact that 80% of comparisons is decided upon within a time investment of less than 6 min, and 50% of the comparisons was completed in a minute and a half or less. Given the fact that time is not normally distributed, the median time investment of 1 min 30 s is a better indicator for the time investment per comparison.

There were some differences between the assessors regarding this time. Of the four assessors, the second had the lowest median time (see Table 8.1), while the fourth had the largest median time per comparison. For the different assessors, the percentage of comparisons done within 1.5 min ranged from 40.6% to 57.4%. The comparisons done in under 6 min ranged from 57.4% to 92.9%.

The final SSR of the briefing assessment using D-PAC was 0.71. The analyses of the SSR evolution showed that the SSR hardly got better after every briefing being judged ten times by the set of assessors (i.e. ten rounds). Therefore, a SSR limit was reached at a value of about 0.70 (see Fig. 8.3). Making (a lot) more comparisons would not have increased the reliability in this assessment. Taking the mean time per comparison, an investment of ten rounds equals a total time for the four assessors of 10.5 h (i.e. 7 min 30 s per briefing paper).

The assessment resulted in a rank order of 84 briefings (see Fig. 8.4). Each of the students needed to receive an individual grade for his/her briefing. In order to mark each briefing, the assessors discussed the upper and the bottom briefing to grade them. The upper briefing received a score of 18 at a 20-point scale and the bottom briefing a score of 8. Subsequently, the intermediate briefings were graded using the rank order out of the CJ assessment. Twenty different grades were given depending on bending points in the rank order. Central briefings with a high overlap in the rank order received the same grade.

Table 8.1 Time needed broken down by assessor

Assessor	No. of comparisons	Median time	% of comparisons completed in under 1.5 min	% of comparisons completed in under 6 min
1	210	1.47	50.5	92.9
2	101	0.78	57.4	76.2
3	210	1.62	47.1	83.3
4	101	3.48	40.6	57.4

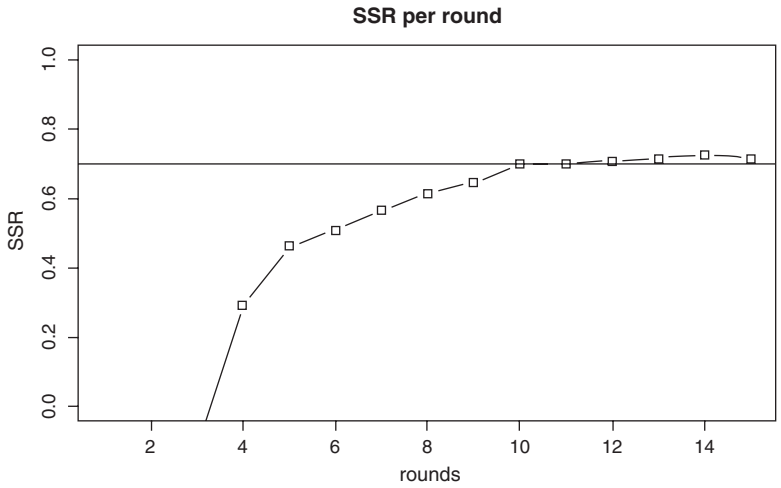

Fig. 8.3 The evolution in reliability per round of comparisons

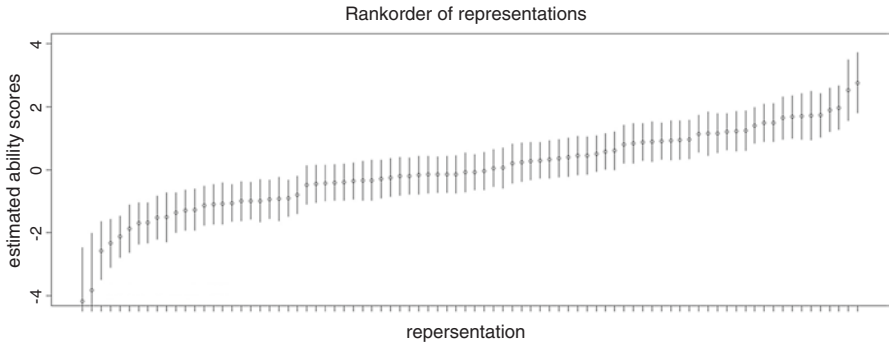

Fig. 8.4 The rank order of briefing papers

8.5 Discussion

The feedback from the participants on the simulation exercise described in this chapter was collected before they received their individual grade. Therefore, although the assessment has been overly positive, it cannot be interpreted as a feedback also on the grades received through this assessment framework.[11] Yet there are some encouraging indications.

Firstly, there is no evidence of a bias towards the most important roles. Whereas participants with most visible and importance roles have performed better than average in the "participation" component of the simulation exercise, such an "advantage" disappears when one looks at the overall grades resulting also from the individual written assignments. At one level, this finding is not surprising. Participants that are more exposed because of their demanding roles have also heavier workload, hence less time to work on the other assignments (i.e. the briefing).

Secondly, as a result of our assessment framework, we ended up with a great variety of grades, ranging from the minimum passing mark (11 out of 20) until very good ones (best students received 17 out of 20). The range and distribution of grades put the simulation game in line with the other courses of the same department.

Thirdly, the comparative judgement method allowed assessing the briefings of the simulation exercise in a reliable yet time-efficient manner. Differences in severity between assessors and within one assessor over time were filtered out, and the final rank order represents the assessors' shared view on the quality of the briefings. The assessors appreciated the user-friendliness and fairness of the tool: compared to the position papers, which were assessed one by one by each assessor, the assessment of the briefings through CJ was felt and described by the assessors as more structured, clearer, more reliable and easier. More than in the time saved, which for some assessors was negligible, the key benefits were identified in the greater confidence they had in the results obtained.

Regarding the element time, all judges completed at least 40% of the comparisons in less than 1.5 min. This appears very fast but could be explained by two elements. First, the judges made notes per briefing paper. When briefing papers reappeared, judgement was likely based on these notes, hence speeding up the decision time.

Second, with 84 papers, an assessor has normally seen all papers once after approximately 42 comparisons. There were thus a high number of comparisons for which the briefing papers had already been seen before and thus made a note on. Independent samples t-test per assessor confirm that for three out of four assessors, the time investment for the first 42 comparisons is significantly higher than for the other comparisons (either 168 or 59). Further research should focus on possible validity issues in using judgements based on notes rather than on rereading the briefing papers.

[11] In fact it is standard practice that a course is assessed before and irrespective of how students have been graded.

Remarkably, the reliability (SSR) showed an upper limit at around 0.70. This is to our knowledge the first time that this occurred in CJ assessments. Usually a reliability limit is only reached with SSR values above 0.80 or higher. Examining what differentiates this study from other studies, the small number of assessors combined with a high number of products stands out. Past CJ studies report nine or more assessors involved in the assessment (Bramley 2015). One study on chemistry products did however reach a reliability level of 0.87 with five judges (McMahon and Jones 2015). Further research should investigate the impact of a small number of judges on reliability in CJ and more specifically on the effect of a judge with varying conception of the competence compared to the other judges. Examining whether any of the four assessors deviated from the group consensus regarding what consists a good briefing paper did not reveal any significant results, but it remains unclear to what extent this index is impacted by the number of assessors. Another contribution of the case of assessing briefing notes to the current field of CJ research is how the rank order is translated into student grades. Although some studies (e.g. Jones et al. 2015) briefly describe that rank orders can be translated into grades, limited studies (e.g. McMahon and Jones 2015) made efforts to describe the process of grading the rank order of a CJ assessment. Similar to what is described in the study of McMahon and Jones (2015), a team meeting was organised to discuss the grade boundaries of the rank order in the current case of assessing briefing notes. However, the similarity of the further process (i.e. grading the intermediate products) remains unclear. For example, McMahon and Jones (2015) do not describe whether they used the (a) position of products, (b) the ability scores of the rank order and (c) bending points in the rank order (similar to our assessor team) or another strategy to ascribe grades to the products. In order to get insight in the value and translation of CJ assessments in practice, more insights are needed into how the rank order can be (and is) used to grade products or to simply identify benchmarks to mark whether or not products provide enough evidence to conclude the candidate possesses the competence under assessment. It is strongly recommended that future research addresses these issues and generates insights into the use of CJ rank orders to decide on grades or sufficient possession of a competence. Given the underlying CJ assumptions, the consensus among assessors in this regard cannot be overlooked. However, starting from this consensus, this question can result in multiple solutions (e.g. discussing a single pass/fail benchmark, using the ability scores to grade products between grade boundaries, insert pre-graded products as benchmarks in the assessment or use an already graded rank order as a starting point for a new assessment). The challenge for future research is to identify methods that are valid considering the CJ assumptions and efficiency keeping in mind the user-friendliness of CJ in practice.

In the present study, the focus was on reliably and efficiently grading students' written products. When students came to see one of the assessors for feedback, the briefing was read anew and the notes written during the judging process were used to inform the feedback conversation. This is an efficient approach if only a limited percentage of students asks for feedback. If the purpose is to provide feedback to most or all students, it may be more efficient to write the feedback while judging the briefing papers. The D-PAC tool includes this feature: for each of the two products, strong

points and weak points can be noted. If an assessor sees this product anew, these notes can be modified. In the feedback presented to the student, the rank order can be included next to feedback commentaries from the anonymised assessors. Linked to this, it would be worthwhile for future research to examine the impact of this feedback. Is it more effective for learning to present both the rank order and the commentaries, or only the latter? Moreover, how does this feedback impact students' self-efficacy, certainly for those with products situated at the lower end of the rank order? Such insight would increase CJ's potential as a formative assessment method.

8.6 Conclusion

Are there any insights from the experience described in this chapter that could contribute to the assessment of simulations as a whole? For one thing, comparison between the quality of the briefings at the end of the simulation and the quality of the position papers in its early phases should show some net learning gains that the simulation has generated. A successful simulation should bring a higher level of sophistication in understanding the issues at stake than the level attained by the simple review of the official documents or press articles. It should also reveal enhanced analytical skills in assessing the situation, as well as in identifying possible solutions. Each briefing will be different, not only because participants have different skills but also because it will inevitably reflect the reality as perceived by each participant.

Not only does the briefings' quality (particularly if compared to the position papers) say something meaningful about the benefits brought about by a simulation exercise; it also enables comparisons across simulation games, so that they can be ranked according to how beneficial they have been in stimulating participants to produce quality deliverables.

Of course, although these comparisons will be based on numerical marks, the assessment of the simulation and the comparison across simulations are not merely statistical exercises. Many factors should be taken into account even in very similar simulation exercises, not least the quality of the participants! The quality of the position papers will be very different even if the simulation is based on the same topic but the students are different. The quality of the briefing papers will have to be assessed relatively to the position papers precisely in order to factor in the different points of departure. Yet, this comparison allows determining that, between two simulation exercises, one has been more beneficial than the other (measured as the delta between the average marks given to students for their briefings and the position papers), even if its overall quality may be "lower" in terms of average marks given to students for the same individual assignments.

Altogether, this study has delivered interesting insights for both practitioners and researchers. Comparative judgement has proven a reliable and efficient assessment method in the context of simulations. Using CJ in simulation assessment has pro-

vided opportunities to learn and investigate how the method can contribute to translating the quality of products into (necessary) grades. Therefore, CJ is promising when it comes to finding a balance between achieving sufficient reliability and grading in an efficient manner.

References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives, 9*(1), 95–104.

Anshel, M. H., Kang, M., & Jubenville, C. (2013). Sources of acute sport stress scale for sports officials: Rasch calibration. *Psychology of Sport and Exercise, 14*(3), 362–370. https://doi.org/10.1016/j.psychsport.2012.12.003

Baranowski, M., & Weir, K. (2015). Political simulations: What we know, what we think we know, and what we still need to know. *Journal of Political Science Education, 11*(4), 391–403. https://doi.org/10.1080/15512169.2015.1065748

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607

Bramley, T. (2007). Paired comparison methods. In J. B. P. Newton, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–294). London: QCA.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment* (Cambridge Assessment Research Report). Cambridge: Cambridge Assessment. http://www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf. Accessed 01 Dec 2016.

Chin, J., Dukes, R., & Gamson, W. (2009). Assessment in simulation and gaming: A review of the last 40 years. *Simulation & Gaming, 40*(4), 553–568. https://doi.org/10.1177/1046878109332955

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher, 37*(2), 1–19. https://doi.org/10.1007/BF03216919

Heldsinger, S., & Humphry, S. (2013). Using calibrated exemplars in the teacher-assessment of writing: An empirical study. *Educational Research, 55*(3), 219–235. https://doi.org/10.1080/00131881.2013.825159

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974

Jones, I., Inglis, M., Gilmore, C. K., & Hodgen, J. (2013). *Measuring conceptual understanding: The case of fractions.* Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/handle/2134/12828. Accessed 1 Dec 2016.

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education, 13*(1), 151–177. https://doi.org/10.1007/s10763-013-9497-6

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

McMahon, S., & Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice, 22*(3), 368–389. https://doi.org/10.1080/0969594X.2014.978839

Obendorf, S., & Randerson, C. (2013). Evaluating the Model United Nations: Diplomatic simulation as assessed undergraduate coursework. *European Political Science, 12*(3), 350–364. https://doi.org/10.1057/eps.2013.13

Perchoc, P. (2016). Les simulations européennes. Généalogie d'une adaptation au Collège d'Europe. *Politique Européenne, 2016*(2), 58–82.

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

Raiser, S., Schneider, A., & Warkalla, B. (2015). Simulating Europe: Choosing the right learning objectives for simulation games. *European Political Science, 14*(3), 228–240. https://doi.org/10.1057/eps.2015.20

Raymond, C. (2010). Do role-playing simulations generate measurable and meaningful outcomes? A simulation's effect on exam scores and teaching evaluations. *International Studies Perspectives, 11*(1), 51–60. https://doi.org/10.1111/j.1528-3585.2009.00392.x

Raymond, C., & Usherwood, S. (2013). Assessment in simulations. *Journal of Political Science Education, 9*(2), 157–167. https://doi.org/10.1080/15512169.2013.770984

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

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286.

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2016, November 3–5). *Comparative judgement and scale separation reliability: Yes, but what does it mean?* Paper presented at the 17th annual conference Association for Educational Assessment Europe. Limassol: Cyprus.

Whitehouse, C. (2012). *Testing the validity of judgements about geography essays using the adaptive comparative judgement method*. Manchester: AQA Centre for Education Research and Policy. https://cerp.aqa.org.uk/research-library/testing-validity-judgements-using-adaptive-comparative-judgement-method. Accessed 01 Dec 2016

Whitehouse, C., & Pollitt, A. (2012). *Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment*. Manchester: AQA Centre for Education Research and Policy. https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_CW_20062012_2.pdf. Accessed 01 Dec 2016