Dr Marcin Smolik
English Institute
Maria Curie-Sklodowska University
Lublin, Poland

Does using discussion as a score-resolution method in a speaking test improve the quality of operational scores?

Introduction

A shift in the assessment of language proficiency from discreet, paper-and-pencil grammar tests to various kinds of performance-based evaluations (writing and speaking exams) has resulted in the necessity to resort to human judgement in the assessment of examinee linguistic abilities. While this shift may be claimed to have brought with it a number of benefits for the very broadly understood language learning/teaching process, the use of raters has brought about, or at least brought to the fore, a number of problems of its own, most notably the problems inherent in the subjectivity of human judgement. Despite numerous efforts on the part of assessment programs and educational researchers to identify features contributing to rater subjectivity, both are now well aware that this subjectivity is simply impossible to *eradicate* and – pragmatically speaking – may only be minimised through the adoption of appropriate *a priori* and *a posteriori* (pre- and post-exam) procedures. The former include such steps as careful rater selection, rater induction and training, and the use of rating scales. The latter, on the other hand, involve such steps as double rating of examinee performances as well as employing appropriate score resolution and/or adjustment procedures, conditional upon the localities of the assessment context.

The present paper takes as its focus one of these post-exam procedures, namely, the use of discussion as a score-resolution method. The need to employ some method of score resolution occurs when two (or more) raters of the same linguistic performance (e.g., an essay or a speaking exam) report different, or discrepant, scores, and only one final score (operational score) is to be reported to the examinee and the public at large. This scenario is typical of educational contexts but is not universal (e.g., it is not the case in figure skating and many other sport competitions). A number of score resolution methods are available and they are all briefly discussed in a later section. What must be stated at this point, however, is that the choice of one method over another is not solely a pragmatic decision but it also carries with it serious ethical considerations. While issues such as practicality and personnel/fiscal feasibility are undoubtedly of essence to any assessment program, it must not be forgotten that the primary purpose of score resolution is achieving overall improvement in score reliability and, first and foremost, the validity of inferences based on resolved scores (Messick, 1989). Great care must thus be taken to choose such a score resolution method which will be both practical and yet – at the same time – its use will not result in scores which put a large number of examinees at a disadvantage. The present paper examines the extent to which discussion as a score resolution method leads to an improvement in the quality of the operational score in the context of a school-leaving high-stakes speaking exam in English.

In the remaining part of the paper I will first briefly discuss four score resolution methods. I will then move on to present the present the assessment context of the study, as well as the study itself. Following this, the paper will present the results of the study and discuss some implications the findings have for the exam as such, as well as for any other exam boards contemplating the use of discussion as a score resolution method.

Overview of methods to resolve score discrepancies

What all score resolution methods have in common is the practice of subjecting an essay or a spoken performance to a second stage of scoring, referred to as adjudication, arbitration, or moderation (Johnson *et al*., 2005). Exactly which productions are re-scored is contingent upon the definition of impermissible discrepancy between two (or more) scores assumed by the exam board. Some of them only re-rate a performance if the discrepancy exceeds a pre-set value, others refuse to accept any disagreement whatsoever.

Exam boards have developed and adopted a number of methods aiding in score resolution. Johnson *et al*. (2005) identify four score resolution models: (1) the *tertium quid* model, (2) the expert judgement model, (3) the parity model, and (4) the discussion model. These models are briefly discussed next.

The characteristic common to the *tertium quid*, expert judgement and parity models is the procedure of employing a third rater to act as an adjudicator. This third rater completes a blind review of a performance (essay, recorded speech) and provides a rating. How the adjudicator's rating and the initial two ratings are now handled distinguishes one model from the others.

In the *tertium quid* model, the operational ratings may be formed in one of the following ways: (1) the adjudicator's rating and the closest original rating are summed and averaged, the other original rating is discarded; (2) the adjudicator selects the operational rating closer to his / her own, and doubles it; (3) the adjudicator moves one of the original ratings up or down. The assumption underlying the use of the adjudicator is that he / she is an expert and thus understands the rating scale better, and applies it more consistently.

Even more trust is placed in the level of expertise of the adjudicator in the expert judgement model, in which the expert's score replaces both original ratings and is the score reported to the test taker and the general public. Such blind trust in the virtual infallibility of the expert may be viewed as problematic. For this reason, the parity model treats all three scores equally, by either summing or averaging them.

The goal of discussion is for the two (or more) raters to arrive at a consensus score, which may entail the necessity to change some of their initial convictions about the performance level reflected in the test taker's response. Reconciling the differences of opinion should not be based on one rater simply surrendering to the overpowering will of another, unless, of course, that rater recognises the superior expertise of another rater, and the acceptance of the other rater's score is not a result of submission, fear, conformism, or indifference, but a conscious choice to defer to the more experienced rater's expertise. In all other cases, discussion must involve an exchange of viewpoints, providing examples for one another to corroborate one's opinion, challenging assertions by providing counterexamples.

The power of discussion has been recognised in the process of writing tests. Davidson and Lynch (2002: 98) argue that "the best tests are not created in isolation by individuals, but are the result of the collaborative effort of a group of people." As to the usefulness of discussion as a score resolution method, there appears to be less consensus. Clauser *et al*. (1999: 42) observe that "discussion of discrepancy in ratings has minimal impact on the precision of the resulting scores" but also admit that the usefulness of discussion may depend on the assessment context. Johnson *et al*. (2005) make a suggestion very much to the same effect. However, while Johnson *et al*. (2005: 141) see discussion as "a viable resolution method when stakes in the assessment are high," Clauser *et al.* (1999: 42) believe that a discussion of discrepancies may prove useful "when the skill under assessment is relatively elementary."

The relative usefulness of discussion as a score resolution method is not the only problematic issue. Another one pertains to the economic feasibility of the use of discussion to

improve test reliability. Besides, there is always the risk that one rater, not necessarily an expert, might dominate discussion and impose his / her opinion on others. This is maybe the reason why Johnson *et al*. (2005: 143) argue that "the raters who are to use discussion should [demonstrate] not only an understanding of the particular rubric but also should [demonstrate] an understanding of the role that discussion plays in the scoring process."

Although the literature dealing with score resolution methods is rather scarce, some initial studies have begun to appear, most of them dealing with rating *essays* using both holistic and analytic scales, with discrepancies resolved using various models. Thus, Johnson *et al*. (2000) investigated the usefulness of expert, parity and tertium quid models for resolving differences on an analytically-scored essay. Johnson *et al.* (2001) and Johnson *et al.* (2003) looked at holistically-scored essay rating discrepancies resolved using the parity and *tertium quid* models (2003), or all four models (2001). Finally, Johnson *et al*. (2005) compared averaging and discussion on essays scored using both a holistic and an analytic scale. Myford and Wolfe (2002) compared the TSE (Test of Spoken English) discrepancy resolution method of third-rater adjudication with identifying suspect profiles of ratings using FACETS (a Rasch-based computer programme; Linacre 2005, 2006). Their analysis identified some profiles of rating as suspect that would not have been so identified using the typical TSE discrepancy resolution criteria.

What these initial studies seem to suggest is that the resolution methods, just like raters themselves, are not interchangeable; the choice of the resolution method might affect the reliability and validity of operational scores. From the point of view of validity, such knowledge cannot be trivialized. If the interpretations of test takers' performances are likely to vary according to the resolution method, then it is only fair to expect that exam boards will provide evidence to the effect that the score resolution method adopted allows for the highest level of rater reliability or dependability (for criterion-referenced testing), as evidenced by limiting the number of misclassifications of the same performance as either a pass or a fail. Otherwise, the trust which we can place in the validity of test score interpretations is limited, and the uses that are made of such test scores of a highly questionable nature.

This necessarily brief literature review also shows a dearth of score resolution studies carried out in the context of speaking exams. One obvious reason for this lack is the very nature of such exams: most of them are conducted and assessed 'live' which basically rules out the possibility of any score resolution even in cases where discrepancies appear (the scores may be later averaged though). The study in this paper is an attempt at addressing the above identified lacuna.

Outline of assessment context

The study described in this paper was carried out in the context of the 'nowa matura' (*new matura*, henceforth NM), a secondary school-leaving examination introduced into the Polish educational system in its final form in 2005.[1] It is a series of non-obligatory high-stakes exams prepared centrally for the whole country by the CKE (Central Examination Board), taken by all students wishing to go into higher education as the NM has replaced university entrance examinations (the rationale behind introducing the examination is discussed in Eckes *et al.*, 2005 and, in much more detail, in Smolik, 2007).

The NM consists of two parts: **external** (written exams marked by trained examiners outside the school) and **internal** (spoken exams marked by both trained and untrained raters in school). With the exception of spoken Polish and spoken ethnic minority languages, all exams are available at two levels: *basic* and *extended*. Candidates are obliged to take a

[1] The exam has undergone a number of changes since 2005 and further changes are expected in 2013.

minimum of three written (Polish, a foreign language, and a subject of student's choice) and two spoken exams (Polish and a foreign language); they may also decide to take 1 – 3 additional exams at the extended level. The present study is a small portion of a comprehensive inquiry made by the author (Smolik, 2007) into the scoring validity (Weir, 2005) of the NM speaking exam in English at the basic level (henceforth, 'the EXAM').

The EXAM is a standards-based test taken by 18-19 year-olds who differ greatly with respect to how long they have been learning English as a foreign language, thus constituting an extremely uneven group level-wise, essentially spanning the whole proficiency continuum. With regard to the administration, the EXAM consists of two tasks and lasts up to fifteen minutes, with five minutes of preparation time and a maximum of ten minutes for the exam *per se*. As far as the scoring process is concerned, at the time of the study (2006) the EXAM was scored using two kinds of scales: a task-specific scale (analytic) and a language abilities scale (holistic). Each examinee was rated by a three-person team consisting of a chair, an interlocutor and one more rater. All team members first scored examinee performances independently and after scoring had completed, all team members discussed discrepancies in scores with the aim of reaching consensus on the operational score.

It might be hypothesised that discussion was adopted as a score resolution method in the EXAM for the following three reasons. Firstly, with the NM constituting a big enough departure from tradition within the largely knowledge-regurgitation-oriented Polish educational context, discussion may have been seen by the EXAM developers as upholding procedures which were well established and generally accepted by teachers and school administrators. Secondly, since – for reasons to do with the necessity to examine over 200,000 candidates within one month – the EXAM may be administered and rated by both trained and untrained teachers with the proviso that the chair in each team always be trained,[2] discussion may have been chosen as a precaution against performance scores being unduly influenced by ratings awarded by unqualified individuals. There may also have been hope that discussion would serve an educational purpose, a form of indirect instruction geared at untrained teachers. While the effectiveness of this strategy has been shown to be largely questionable (Smolik, 2007), it does nevertheless point to the fact that some care had been taken *a priori* by EXAM developers to prevent inconsistencies from occurring in the testing and rating procedure. Thirdly, of the four score resolution methods outlined above, discussion may have seemed the only viable one considering the fact that operational scores must be reported to candidates on the day of the exam.

The choice of discussion as a score resolution method is not, in and of itself, inappropriate. However, in the case of the EXAM the decision does not appear to have any grounding in research. I am not aware of any studies, published or not, which would have been conducted with the view to establishing the variability in the rank order of students' scores, their mean scores, or pass-fail decisions depending on the resolution method adopted, and as Johnson *et al.* (2003: 317) usefully caution, "the interpretations of students' performance are likely to vary according to the resolution method." Considering the stakes of the exam such considerations are all but insignificant.

[2] If discrepancies occur between marks assigned by individual raters and they cannot reach a consensus, the final decision always rests with the chair who, it is believed, as a trained examiner, simply 'knows better'. As my research has shown (Smolik, 2007), the differences between trained and untrained raters are largely marginal, which may be due to the fact that in the case of NM speaking exams, once trained, the certified raters are not required to undergo any additional re-training or moderation sessions prior to every EXAM administration, which is common practice in many examination boards (e.g., Cambridge ESOL).

The study: Research questions

The data reported here form part of a larger investigation into the scoring validity of the EXAM (Smolik, 2007), a study which I carried out in 2006 and 2007. In what follows, I will only provide research design details and sample characteristics which are of relevance to the immediate topic of the paper.

The present paper aims at providing answers to the following research questions:

(1) What is the nature of discussion as a score resolution method in the EXAM?
(2) Are raters equally engaged in the resolution process, or does the use of discussion as a form of resolution allow the opportunity for one rater to dominate, or defer, to others?
(3) Can discussion be claimed to work effectively as a score resolution method in the EXAM?

The study: Participants

Participants were 20 secondary school teachers of English (all volunteers), 12 of whom were certified EXAM examiners and 8 were not. Eighteen teachers were female and 2 were male. Most of them were between 26 and 35 years of age, with an average teaching experience of 11.5 years and 3 years for trained and untrained examiners, respectively. All teachers had had prior experience with the EXAM. One of the certified examiners had been asked to choose three students of hers to participate in mock exams. These were videorecorded and digitalised for further use.

The study: Research design

With the exception of the interviewing teacher, the remaining 19 participants were asked to form examining teams, i.e., get into pairs with the proviso that there be at least one certified examiner in each pair. The certified examiner would act as chair, the second teacher in each pair would be the team member while the third team member (the interlocutor) in each case was the teacher who conducted the three mock exams. Eventually, 10 pairs were formed (it was necessary to form two pairs consisting of two examiners, as well as have one examiner 'chair' the discussion twice).

The ten teams watched the three videorecorded mock exams, and each meeting progressed according to the following agenda. First, the teachers were briefed on the procedures, following which they watched the three recordings, assigning marks individually as they watched using the rating scale routinely used in the EXAM. Once a recording finished, the team was asked to work together and reach a consensus on the final score. The interlocutor had been instructed to 'stick' to the scores she had assigned immediately upon the completion of the recordings and to offer the same explanation for the scores in the process of discussion with each team. The discussions were audio-taped with the subjects' permission.

As instructed at the beginning of each meeting, the person (self-) appointed chair was responsible for observing the pre-arranged time limit of about 5 minutes per recording. The time limit was set in result of observations made in a pilot study of the methodology, as well as in order to reflect genuine time limits routinely applied during live EXAMs. Where compromise could not be worked out, the final decision concerning a mark rested with the chair. Altogether, 30 (10 x 3) team discussions (TDs) were collected (audiotaped), and these were later transcribed for further analyses.

Additionally, in order to evaluate the quality of ratings assigned by the teams, it was considered necessary to obtain an assessment of each of the three mock exams which could serve as a benchmark against which wellgrounded comparisons could be made. Such model ratings were provided by two experts, both Central Examination Board employees.

The study: Data analysis

Upon the completion of the study, the databank consisted of (a) scores assigned to the three videotaped mock exams, assigned both by individual team members as well as the teams together, and (b) 30 transcripts of discussions. The former were subjected to quantitative analyses, utilising descriptive statistics and correlations, the latter were subjected to a qualitative analysis geared at identifying the salient features.

The study: Results and discussion

The findings of the study are presented as answers to the three research questions posed above.

Research question (1)

The procedure whereby members within a team first score examinee performances independently and after scoring is completed all team members discuss discrepancies in scores with the aim of reaching consensus on the operational score, was uniformly adopted by all the teams in the study. Upon finishing watching a videotaped performance, the chair initiated a discussion by directing other team members' attention to the first item in the EXAM set. The discussion then proceeded on an item-by-item, sub-task-by-sub-task basis, with the chair noting down the final score in each item/(sub-)task/criterion in the protocol, just as it would be done in the live EXAM. Eleven of the thirty recorded discussions concluded with the chair reviewing the scores assigned, typically followed by a comment on the examinee passing or failing the exam, which might reflect a practice gained through participation in live exam proceedings.

While the overall structure of the discussion was virtually identical across all teams, considerable variation was observed with respect to the duration of particular discussions. Although the teams were instructed prior to the recordings that the score resolution process should last approximately five minutes, twelve discussions were longer than six minutes, the longest being ten and a half minutes. On average, a discussion took little below six minutes (5'53'').

Although the relevance of this finding might appear marginal, it is not necessarily so. First of all, it might indicate that the discrepancies in the scores assigned by individual raters were so marked in some cases that any consensus could only be reached following a long discussion and a considerable number of (forced?) compromises, especially if two or all team members held strongly to their initial positions. Secondly, the inability to reach consensus quickly might indicate that some chairs lack the skills that are required of one in charge of discussion proceedings. Finally, considering the real-time constraints of the EXAM, a balance must be found between the quality of the discussion and its duration. It could be argued that a more thorough analysis of an examinee's performance, providing arguments and counterarguments in support of a particular score *might* result in more accurate (and hence valid) ratings. However, the process cannot – for practical reasons – last infinitely; it may lead to problems with test administration (if, for instance, many students are examined on the same day) and, much more seriously, test score quality (overlong discussions extend the duration of an exam session, potentially leading to rater fatigue, which might exert an adverse impact upon the scores the rater awards).

With regard to the lines along which the discussion progressed in the individual items/(sub-)tasks/criteria, a number of 'strategies' could be distinguished, depending on whether all three members awarded an identical score individually, or if there were some discrepancies between the individual raters.

Thus, when there was absolute agreement among the team members with respect to a score, the discussion often turned into a mere recitation of the scores, with no explanation provided, e.g.:

Quote 1: [3]
e8: the first situation, the first message
n2: well, I've got one
e0: yes
e8: one, uhum

At times, the information about the number of points assigned individually by raters was supplemented in the discussion by one or two team members with a relevant quote from an examinee's performance, or with its Polish translation, e.g.:

Quote 2:
e4: OK, now moving on to the third, yyhm, why, yyy, OK, here
e10: the first piece of information, I've got one
e4: I've got one as well
e0: one
e4: one, 'I was in English', yes
e10: there was a reason too, 'I must learn'

When there were some discrepancies in the scores assigned by raters individually, the discussion proceeded along slightly more varied lines, yielding even more varied outcomes. In all cases, however, the raters engaged in the process of mutually reviewing student performances, typically with two raters providing arguments to convince the third rater to the rightness of their position. In most cases, the process was a fairly smooth one, with the rater who awarded the discrepant score acquiescing to the opinion of the majority, e.g.:

Quote 3:
e7: ehm, fascinating features
e0: there were those
e7: there were, 'beautiful', 'calm' and he was in Africa
e0: 'beautiful', 'calm' and he was in Africa
n7: well, I didn't give a point here because, well, he did say, at the beginning, that, that he met a 'beautiful man', but these features, he said, he only gave 'calm', and so I marked here that, that there was only one feature, but if we accept the beginning, what, what he said, that he was beautiful, that 'beautiful man', then, well, I agree but I don't think that being in Africa is a feature
e7: well, it's not a feature, I agree
n7: and these were supposed to be features of a newly-met person
e7: mhm, he said that he was 'beautiful' and 'calm', I'd accept this
n7: well, in this case, yes, yes, if that's how we see it, then OK

At times, achieving consensus by persuading one rater to adopt the opinion held by the other two proved to be a lengthy and laborious process, with raters resorting to arguments from an ever-widening circle of options. The quote below, not altogether exceptional with regard to its length (over two minutes of conversation), illustrates this phenomenon very well.

Quote 4:
e8: the third message?
e0: a point
e8: a point
n8: the third? I've got zero
e8: because?
n8: because if he said, well, the question is, how did this meeting influence your journey?, and he said that it didn't, he said, 'no', he answered the question?
e8: yes

[3] The TD transcripts have been translated from Polish by the author. In the transcripts, rater codes are printed using different font type: bolded for chair, italicised for interlocutor and normal for team member.

e0: uhum
n8: and then he said that he talked to his grandmother, and he was only going to his grandmother
e8: but, but he was traveling with her
n8: no
e0: no, he visited her later
n8: no, he was traveling with this person
e0: to his grandmother
n8: who had pink hair, he was traveling to his grandmother
e8: to his grandmother
n8: and then he talked about her with the grandmother
e0: but since, since here we've got how this meeting influenced the journey, if it didn't
n8: it didn't influence
e0: then the message has been conveyed, it's not stated anywhere that it was supposed to influence it in any way
n8: but it says and DESCRIBE how it influenced, then, well, I think that he should say something here about this influence
e8: and if, no, well, he said it DIDN'T influence, I mean, no, this is, well, because it is presum-, presum-, y, y, y, this is an instruction with a presumption, and if he doesn't agree with this, and he also said, he referred to this because he said that
e0: but then he
e8: BUT I talked, so he says, no, nothing happened, but I know this is too short, just like the other one so I'll give him here
e0: well, he, he, he, he, with this extra message that later he talked with the grandmother about it, this explains that he knows what this is all about, really
e8: yes
e0: that it wasn't, well, accidental
e8: it seems so to me as well, that, that he came, that it didn't change anything in his journey, so it didn't change anything but on the other hand
n8: it may be so, I agree with what you're saying but I think that they should stick to the scripts and not, not change them as they please, so if it was supposed to influence the journey in one way or another, he should say that it did
e8: well, because it is so that, for instance I know from experience that if someone says, what did you have in the bag? and someone says, I didn't have anything
e0: exactly
e8: and this is a good answer, acceptable
e0: yes
e8: I don't know, I'm not going to dig my heels in on this one but it seems to me that we could accept this answer because he did deal with the situation and also explained that he knew what was going on, that later he talked
n8: uhum
e8: because for instance, the previous one [=student], it is as if
n8: mhm, OK, I understand, <u>you've convinced me</u>
e8: right, he didn't assume, I mean, it seems to me that he didn't assume anything had to happen
e0: of course
e8: she got on, she got off, he watched, and by the way he talked with his grandmother about it
n8: <u>all right, you've convinced me</u>
e8: that's what it seems to me
n8: let it be
e0: uhum

While a discussion of this kind may be a valuable learning experience for all raters, from the point of view of exam practicality, such practices are questionable and, arguably, should be discouraged. Live exams are not the right time and place for such debates, which serve the function of examiner training. That the training is necessary – the quote proves beyond any doubt, only it should be done *prior* to the exam, during a moderation session for all teachers who are going to act as raters, not while the exam is in progress.

In addition to being a training tool, the score resolution discussion was also found to serve a monitoring function, safeguarding against inappropriate scoring decisions due to a failure on one rater's part to notice a particular piece of information in an examinee's discourse in result of fatigue, boredom or a moment's inattention. That discussion *does* perform as a quality check on individual rater's work must be seen as one if its greatest advantages:

Quote 5:
e8: second message?
n2: zero
e0: one
e8: one ((laughs))
n2: second, meaning what?
e8: hem ((??)), it was
n2: the features, well, there were no features, if you ask me, 'he was interes-ted', no, 'I was interested'
e0: but he said, 'calm', 'beautiful'
n2: he did?
e0: yes
e8: yes, at the very beginning
n2: 'beautiful'!, yes, I remember, I remember 'beautiful'
e8: then there was 'calm'
e0: then there was 'calm' and also 'in Africa'
e8: and Africa was as ((???))
n2: all right, I admit, Jeez!

Raters' discussions on the scores for individual criteria / items also varied with respect to their outcomes. In the event that common consensus was reached, the final scoring decision was uncontroversial. However, in those cases in which consensus could not be arrived at, the final decision rested with the chair, who tended to opt for either of the following options. If two team members agreed on a score and the other one did not, and it was impossible to convince him/her to embrace the opinion of the majority, the chair, more often than not, accepted the majority vote as the final one, especially if his/her own voice was in this majority (cf. Quote 6).

Quote 6:
e3: fine, relating events, I have to admit, in the first one, in the first one, I don't have [a score] here
e0: I've got a point
e3: I also accepted the circumstances, 'in supermarkt'
e0: yes
n5: but it's, it's 'Peter is supermarkt'
e0: no, she met him in a supermarket on holiday
n5: I didn't understand it this way
e0: 'I holiday'
n5: what I got was 'I meet Peter is supermarkt'
e3: yes
n5: 'when I holiday', it was, 'when I holiday'
e0: no, 'I holi-', well, OK, there was 'supermarkt', it's true but 'in' or 'is', it's a small difference
n5: aaaaaa
e0: only one little letter
e3: well, she really, here, that's true, but she definitely said something about 'two people'
e0: she was on holiday
e3: she met, just by the way
e0: yes, yes, and his name was Peter
e3: she only talked about this Peter ((laughs)), all right, but she was 'very interesting', and, and, and, and, and there really was 'Peter is supermarkt'
n5: uhum
e3: I don't know, maybe he's super simply
e0: but 'on holiday', so these are circumstances, I'd give a point for this
n5: it wasn't even 'when I'm on holiday', only 'when I holiday', that's what I've got noted down, 'when I holiday'
e0: yes, that's what it was
e3: yes, that's what it was, but I'll accept it, well, 'when I WAS on holiday'
n5: well, OK
e0: well, I accept it too, I'd figure it out if someone told me so
e3: because of 'when'? she didn't say 'I holiday' but 'when I holiday'
n5: all right, jaaa, well, I'll
e3: I'll give her one point
n5: you decide

e3: just like, well, I decide but Magda [=*e0*] supports me here

The situation was less predictable if the one rater at odds with the majority was the chair him/herself. While in some such instances, the chair acquiesced to the voice of the majority:

Quote 7:
e6: reject his proposal of a solution
n9: she rejected, 'no I must learn', something, something
e6: I gave zero here as well
n9: no, she said, 'no, I must learn'
e0: certainly
n9: only I can't remember what next
e0: 'because I exam'
e6: 'I exam'
n9: 'because I have exam'
e0: no, 'because I exam'
e6: no, 'I exam', 'because I exam'
n9: aaaa
e0: but before that she said, 'no'
n9: yes, there was 'no'
e6: all right
e0: there was, there was
n9: 'no, I must learn because I exam'
e0: that's what it was
n9: something like that
e6: let it be, I didn't accept it here, but let it be, OK, there is a reason
n9: there is a reason

it was also common for the chair to exercise his/her authority in such situations and award his/her own score as the final one, sometimes clearly against the will of the other team members:

Quote 8:
e3: now, a question, did he reject his roommate's suggestion?
e0 / n5: he did, he said 'no' (…)
e3: yes but 'reject the suggested solution' for me also means give a reason
e0: there's nothing like that in the instruction (…)
e3: Beata [=n5]?
n5: well, I know from experience that if they reject, say 'no', they are given the point, so
e3: well, it depends
e0: I mean, in this communi-, in everyday communication if someone asked us, hey, listen, yyy, do this and that, then we say, no, and what I suggest is rejected, in everyday communication, this is a conversation with a friend, we can't expect a friend to say, listen, well, I don't entirely agree with what you say, right, because something, something, you simply say, no, or I can't make it
e3: but still, he's supposed to suggest another solution, so how, how can this be a natural conversation?
e0: well, theoretically, it's supposed to resemble one, mhm
e3: well, excuse me, then I'd say, 'no, sorry', so that there is something more than only 'no' or 'yes'
e0: but that's what he said so I think that
e3: I'm sorry ladies but I'll exercise my authority at this moment
e0: do it
e3: it's zero points

Which particular course of action the chair decided to take when there was disagreement appeared to depend on the individual. Of the 10 instances in which the final rating decision was taken by the chair exercising his/her authority against a majority decision, seven were associated with only two examiners. It would appear, then, that the chair's personality characteristics may impact upon the practices he/she adopts. This conclusion must be treated as an indication of a possibility; with only three performances assessed by teams, making broad generalizations had best be avoided. Yet this initial finding, although largely intuitive, seems to echo a similar finding obtained by Dubiecka *et al*. (2006) in the context of

a written competence test for primary school students, who observed that some variance in test scores appeared to be attributable to the person coordinating the rating process who could be considered an equivalent of the chair in the EXAM.

While examining the different paths leading to the final score was an important endeavour, a study of the nature of discussion as a score resolution method would be incomplete without attempting to identify the reasons underlying the differences in the final scores assigned by all the teams, for the differences were sometimes quite pronounced (this is discussed in more detail in research question 2). An analysis of the TD transcripts revealed that the teams struggled with largely similar problems to those that individual raters were found to hurdle over.

Thus, the differences between the final ratings assigned by teams emerged in result of an approach adopted by a team with respect to an interpretation of words such as 'features' or 'circumstances' which appeared in the role-play scripts (cf. Quotes 3, 5 and 6). Working in a team did not appear to help raters work out universally accepted procedures for dealing with examinee responses which were either very short (cf. Quotes 7 and 8) or contained an 'empty set', an answer in which an examinee uses e.g., 'nothing' instead of an object (cf. Quote 4). Teams also struggled with ways of dealing with the cases of unnecessary interlocutor support. The issue virtually appeared to split the teams into two opposing camps, the split clearly proving a lack of clear instructions on the appropriate course of action in such an event. Interestingly, the split was there despite the fact that the interlocutor shouldered the blame and urged all teams to award a point. Quote 9 illustrates the issue:

Quote 9:
e3: how did this meeting influence the journey? what can you say here, ladies?
e0: I didn't mark anything here because I didn't remember back then but it seems to me that, yhm, I mean, he deserves this point, which is my fault, because he didn't squeeze anything out of himself, and I suggested the answer strongly
e3: exactly
e0: one way or another, he deserves the point (…)
e3: well, I awarded a point here as well, unfortunately, thanks to the interlocutor
e0: he should get it, yes
e3: well, what can we do? we have to take the consequences

One consequence of the lack of guidelines for tackling such situations appears to be that teams worked up their own, internal set of guidelines in order to facilitate the rating process. Consider the following quote taken from one team's discussion aimed at reaching consensus on the score for a student's performance on an item. The recording featuring the student (Łukasz) was watched by the teams as the last one in the series and the underlined phrases clearly show that the team members acted according to previously agreed guidelines. It is important to reiterate that no such guidelines were (and are) officially issued by the EXAM developers; they were created by the team while rating was in progress, in response to a perceived gap.

Quote 10:
e10: I mean, keeping to what we did previously, I awarded zero points, because we had there, yhm, describe how this meeting influenced the journey, with the previous conversation, there was this Africa thing there, and here, well, well, there was no influence on *this* specific jour-, journey, that's why
e0: well, there wasn't but he, but he said
e4: but on the journey as such?, that he talked with his grandmother?
e0: I'm not talking about that
e10: 'I talk to my grandmother about it' (…)
e4: well, if we are to keep to what we adopted there, that if an interlocutor helps a bit then (…)

On the one hand, such a course of action must be seen as a positive endeavour for it could be argued to enhance rating consistency *within* a team. On the other hand, however,

such instances of what we could term a 'glocalisation' of the rating criteria may lead to a situation in which every team practically uses idiosyncratic, largely incomparable set of criteria, ruling out any possibility of score comparison. The adverse consequences of such a situation for scoring validity are transparently obvious.

Other problems with the understanding and use of the rating scale experienced by teams included, unsurprisingly, interpreting some vague phrases contained in the descriptors and basing the ratings on criteria-irrelevant information. For instance, one team had the following discussion with regard to 'answering the questions' (one of the EXAM tasks):

Quote 11:
e0: I don't know, he could have said more, really
e4: well, to all intents and purposes, he could have, I've got one point here as well, but
e10: yeah, exactly, this is so
e0: I mean, I don't know, I gave two and
e4: this answer is always the biggest problem
e0: uhum
e4: because it's never clear, how much is 'sufficient', is it?, the first question, I mean, well but
e10: is it enough or not enough

The criteria-irrelevant information attended to by teams were the same ones that I identified in an analysis of individual rater's stimulated recall reports (not discussed in this paper): making comparisons between examinees, considering linguistic ability when rating for conveyance of information, or – on the contrary – making the mark for 'linguistic ability' conditional upon the number of messages conveyed, judging an examinee as 'communicative' or otherwise, considering examinees' personality features, judging an examinee's comprehension abilities. In fact, the accumulation of criterion-irrelevant information heeded by some teams is such that no criterion-relevant information is taken into account when working towards reaching a consensus on a score. Consider the following quote (criterion-irrelevant information is underlined), which, while appearing rather extreme, is really quite representative of the approach exhibited by a number of teams.

Quote 12:
e7: OK, what about language?
e0: one
n7: two
e7: I've got two, I've got two, despite everything, well, yes, despite everything, despite everything two here, it is, let's say, a bit 'stretched', and I'm a bit influenced here by what I heard yesterday [=in a live exam], really, what the exams were like yesterday ((laughs)), although they were much better but
e0: I mean, I, mmm, I gave one because, mostly because he had problems conveying many messages, he wasn't very communicative as far as the content is concerned
e7: I mean I agree, he wasn't very communicative and, well, yes, because he repeated all the messages
e0: gaps in vocabulary, serious ones
e7: uhum
e0: and a complete lack of self-confidence in communication, if you ask me
e7: okay, I mean I would be willing to give those, I mean, he, that's what I say as well, I gave two but it is a 'stretched' two, so I would be willing to lower it
n7: to one
e7: to one, well, okay, fine

The final comment on the nature of discussion as a score resolution method will concern a largely elusive characteristic, namely, the style in which the discussions were conducted. It cannot be overemphasized that any precise categorizations with respect to this issue are impossible, for the evaluations were largely subjective, made by the present researcher on the basis of his own observations as a mute participant and transcriber, and as such of necessarily limited credibility. Nevertheless, the brief discussion is offered in the context of awareness of its limitations.

Generally speaking, the discussions were conducted in a friendly, yet professional manner, although occasionally some raters exhibited an inclination towards caustic comments made in relation to a particular student. Such comments, however, were never snide or cruel. Sporadically, the team members engaged in something that could be interpreted as a kind of 'competition' aimed at providing as distorted – yet still linguistically valid – a version of an examinee's utterance as possible. Such activities may be seen as a way of dealing with anxieties inherent in the rating process, but also as a way of creating something akin to 'corporate group culture', if only temporarily.

All discussions were conducted in a relatively dynamic manner, with raters often speaking simultaneously, which must not be interpreted as 'attempting to shout one another down' for no such cases were observed, although – occasionally – one rater, typically the chair – tended to dominate the discussion. Generally speaking, the chairs tended to adopt one of two styles in conducting the discussion: 'friendly' or 'panel'. The former resembled, to a considerable degree, a discussion among friends, with all team members sharing responsibility for structuring the interaction. The latter was more 'professional' in the sense that the chair first provided his/her own rating with a justification, and then listened to other team members' opinions, after which he/she announced the final decision. The two styles were distinguished not so much by the level of formality of language, however, as by the procedures adopted in speaker nomination, power relations, etc. Without more detailed ethnomethodological studies into the dynamics of team interaction, it is impossible to state with any degree of certainty whether either of the two styles had any influence upon the scores assigned, although the issue of how teamwork is influenced by the degree of familiarity among team members, by individual sympathies and antipathies, is surely one of the EXAM-related issues meriting further investigation.

The examination of the nature of discussion as a score resolution method in the EXAM did not reveal any serious deviations or irregularities across teams with respect to how the discussions were structured and conducted. What was uncovered, however, was that virtually all problems experienced by individual raters when assessing examinee performance were also problematic for the teams, indicating that the problems should best first be remedied at the individual level. The remedy, again, appears to be relatively straightforward and involves the steps well known to the testing community: meticulous set standardization, rating scale validation and a well organized system of examiner training and moderation.

Research question (2)

Implicit in the use of discussion as a score resolution method in the EXAM is the expectation that *all* team members do not merely acquiesce to the most assertively voiced opinion, but are equally involved in the process of arriving at the operational score. Active participation does not, by any means, imply domination, on any team member's part. Even the chair, possessive of the authority to reject the voice of the majority, should take heed of what the other team members have to say and should at all costs avoid a situation in which his/her voice becomes one of an authoritarian regime rather than sound argumentation, reason and counsel.

In order to find out whether the operational scores awarded by teams tended to agree more closely with the scores of one team member (chair, member, interlocutor) than with the scores of other team members, the scores assigned to the three videorecorded performances by individual team members independently (prior to discussion) and the final scores awarded to the same students by teams were tabularised and a comparative analysis was carried out. The analysis was geared towards identifying whose voice, if anyone's, tended to be the most authoritative in cases of discrepancies between the scores assigned by individual raters.

To that end, all possible 'patterns of disagreement' were first established, and they are presented graphically in Fig. 1. In a team consisting of three individuals, four such patterns

are possible. In the first three, two team members agree on a score while the third one assigns a different score (specifically: 1: chair and member *vs.* interlocutor; 2: chair and interlocutor *vs*. member; 3: member and interlocutor *vs.* chair). In the fourth possible pattern, all team members award discrepant ratings.

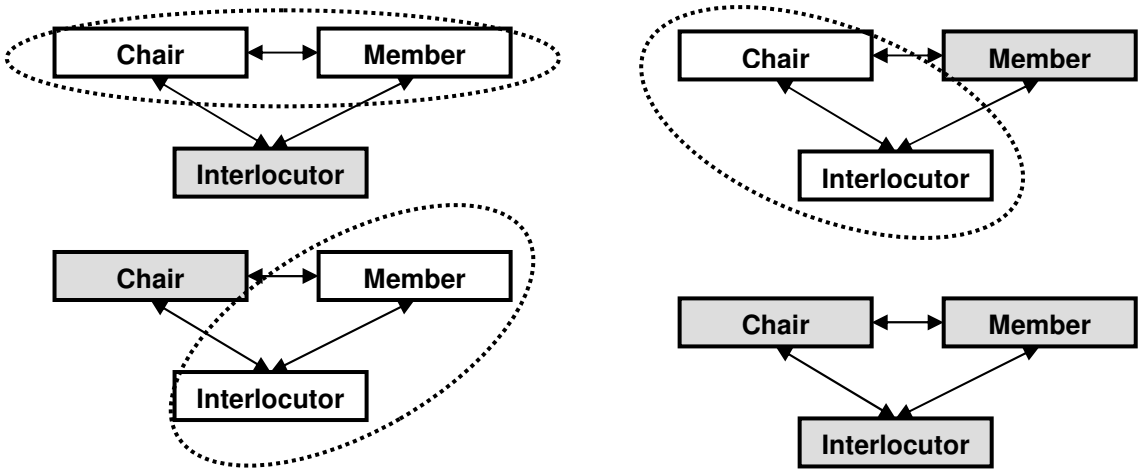

Fig. 1: Possible 'patterns of disagreement' in a group discussion in the EXAM

In order to find out what decisions were actually taken in each of the situations depicted graphically in Fig. 1, the tables containing the scores assigned were scrutinised for all instances of disagreement corresponding to either of the patterns above, and the 'winner' in each disagreement was noted. Altogether, of the 370 absolute agreement opportunities, disagreement among team members was observed in 135 cases (approximately 36%). Fig. 2 presents the results of the analysis of 'winners' in all situations depicted in Fig. 1.

As can be seen in Fig. 2, the final score assigned to an examinee in a particular criterion when the individual raters disagreed was, in a vast majority of cases (74%), the score on which at least two team members agreed. I argued previously that it was often so that the two team members who were in agreement with each other managed to convince the 'dissenting' team member to their opinion. It would thus seem that in the EXAM the final score is, in most cases, the 'dominant' score, i.e., the score assigned by either all three or two team members. Although the divergence in the frequencies with which either side in the particular 'disagreement patterns' was the winner were quite pronounced, a chi-square test was carried out in order to find out whether the differences in the frequencies were not due to chance. The test, as was predicted, reached significance ($\chi^2(2)=46.403$, $p<.000$).

Although the operational rating assigned by teams for the individual items / sub-tasks / criteria was found to be the rating of the majority in most cases, it seemed to me that the discrepancy in the ratio of instances 'won' in each of the patterns in Fig. 2 by the pair as opposed to the one dissenting rater was suspiciously pronounced between patterns marked as A and B in Fig. 2, and pattern C. In this last pattern, the dissenting voice belonged to the chair, who occasionally tended to exercise his/her authority to award his/her rating as the final one. Of course, in some cases the chair managed to convince one or even two other team members to his/her opinion in the course of discussion, and the reported frequency of 24 does not reflect this, being based on the raw scores.

In order to find out whether the chair's ratings were really the most likely to be chosen as the final ratings, I calculated the frequency with which the final rating agreed with each team member's original rating *across* the four patterns depicted in Fig. 2. Although I realised that the obtained frequencies *did not* reflect all the cases in which other team members

actually *modified* their initial ratings, I believed that such an analysis would allow me to detect a certain tendency in the rating data.

The frequency with which the final score agreed with the original score of each team member was as follows: chair – 48% (112 instances), member – 30% (70 instances), and interlocutor – 22% (53 instances). I computed the chi-square goodness-of-fit index to test the null hypothesis that the expected distribution was 33% for each of the team members. The value of the statistics was $\chi^2(2)=11.286$, $p<.004$, indicating that the final scores agreed more frequently with the original scores of one of the raters, which could be taken as indicative of a possibility of rater dominance or deference.

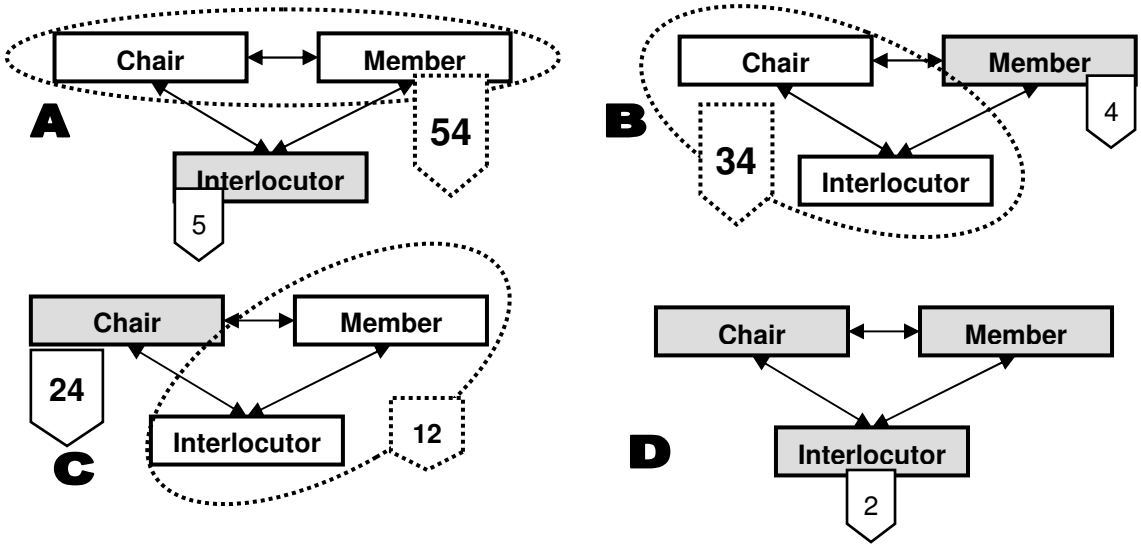

Fig. 2: Frequencies of instances with which either side in 'disagreement patterns' were 'winners'

In order to determine which source was a major contributor to the finding of statistical significance, following Johnson *et al.* (2005: 139) I computed the value of standardised residuals associated with each rater using the following formula: $R=(O:E)/E^{1/2}$, where 'O' is the observed frequency and E is the expected frequency. Values of standardised residuals in excess of |2| may be considered as indicating major sources contributing to the significant chi-square value. The standardised residuals for the individual team members were 3.80 for the chair, -.94 for the member, and -2.85 for the interlocutor, indicating that both the relative frequency with which the operational score agreed with the chair's original score, as well as the relative infrequency with which the operational score agreed with the interlocutor's original score were major contributors to the obtained chi-square value.

This finding may be hypothesised to indicate two things. First of all, the perception of the performance by the interlocutor differs from that of the other team members, hence his/her scores are less likely to agree with theirs, and thus are less often reflected in the operational scores. However, considering the fact that only one interlocutor participated in this stage of the study, this conclusion must be treated with caution.

Secondly, this finding may also indicate that the chair is the dominant force in a team, or that other team members tend to defer to his/her decisions. On the one hand, this may have been caused by the fact that the procedures actually *expect* the chair to have the decisive voice. And under the right circumstances there would be nothing wrong in this situation. This may simply be a reflection of a 'deference effect', whereby a less experienced rater simply recognises the chair's superior expertise, resulting from his/her status as an examiner and

experience, and accepts the chair's scores over his/her own in the belief that they are more accurate. If this was the case, such deference should be welcomed rather than discouraged.

However, as my research (not reported here) showed that there existed no or only minimal differences between examiners and non-examiners with respect to inter- and intra-rater consistency of scoring, the level of inter- and intra-rater agreement, the use of the rating scale, etc., the 'greater expertise' argument above cannot be said to hold. There is nothing in my other findings that would suggest that examiners' ratings are in a substantial way better or more accurate than non-examiners'. Thus the observed dominance / deference effect should not be viewed with heightened enthusiasm, for it may simply indicate that the chair's opinion was believed to be 'right' even if it was not necessarily so.

Without doubt, the fact that both the interlocutor and member sometimes embraced the chair's point of view (and score) in the process of discussion further complicates the attempts to interpret the obtained findings unambiguously. However, I would postulate that there is no ground to believe that the expertise of the examiners is in a considerable way superior to the expertise of (at least some) non-examiners and, hence, *assuming* that the chair's decisions are always right is simply unwarranted. The assumption that 'examiners know/do it better' finds no confirmation in the empirical findings collected in my study. For this reason, the role of the chair in the score resolution discussion, as well as the scope of his/her competences must be subjected to careful scrutiny.

Research question (3)

In order to answer this question some criteria needed to be worked out against which the efficacy of discussion could be judged. For the purposes of the analysis, the following criteria were embraced. Any score resolution method might be claimed to work effectively if its application exhibits the following characteristics:

(1) it is successful in considerably reducing the discrepancies amongst individual raters, so that any discrepancies that are found to exist between ratings assigned to the same performances by different teams are *much smaller* than the discrepancies in the ratings assigned by individual raters;
(2) the final scores assigned by different teams to the same performances are either identical or differ only minimally;
(3) using a different method of score resolution cannot be proved to perform better than the method being investigated with respect to characteristics (1) and (2);
(4) the final scores arrived at using the score resolution being studied correlate better with the scores assigned to examinee performances by experts than scores obtained using a different method of score resolution.

Each of the four characteristics will now be addressed in turn.

Table 1 presents the basic descriptive statistics computed for the aggregate scores assigned to the three students (Ania, Dominik and Łukasz) by individual raters prior to discussion and by the teams. Although the ratings assigned by teams exhibit a smaller range, the difference in range between these ratings and the ratings assigned by individual raters is not very substantial. While individual raters' aggregate scores for the three examinees differ by as many as 6-7 points, the range is reduced to 3-5 points in the case of teams, which is still quite considerable (the maximum score is 20 points). Also, the value of standard deviation was substantially reduced only in the case of team ratings assigned to Ania (a .58 reduction). Hence, it would appear that using discussion as a score resolution method failed to reduce the discrepancies observed at the individual level to an extent which could be considered satisfactory, *at least with respect to the three performances used in the study* (cf. point 1 above).

Table 1: Descriptive statistics for ratings assigned by individual raters and teams

Statistics / Student	Individual scores						Team scores					
	n	*M*	*SD*	*Min*	*Max*	*Range*	*n*	*M*	*SD*	*Min*	*Max*	*Range*
Ania	30	7.53	1.50	5	11	6	10	6.80	.92	6	9	3
Dominik	30	7.73	1.34	4	11	7	10	7.30	1.56	4	9	5
Łukasz	30	14.27	1.74	11	17	6	10	14.60	1.71	12	17	5

A proviso that needs to be made at this point is that the findings reached in this research question must be interpreted with great caution in the face of a relatively limited dataset on which they are based. The conclusions concerning the efficacy of discussion as a score resolution method can only be made safely with reference to the three performances assessed in the study.

Despite the observed discrepancies in the *aggregate* scores assigned by individuals and teams to the same performances, in all fairness, it must be stated that discussion *did* manage to eliminate some differences with respect to the scores assigned in individual items/sub-tasks/criteria, so much so that in many cases the ratings awarded by *all* teams were unanimous.

The sizeable range of team aggregate scores described above would also appear to indicate that the requirement specified in point (2) was not met by the data in the study. Table 2 shows the aggregate scores assigned to each student by all ten teams. As can be seen, in each student's case one score (shaded) appears to 'stand out' in that it is separated from the remaining aggregate points by a missing rating category. For instance, there is not a score of 8 points in Ania's case, thus the scores might be claimed to be attributed to error, at least from the point of view of statistics because it real-life terms we may say with a high degree of certitude that Ania would *not* mind being assessed by Team 7, the score being erroneous or not.

Table 2: Aggregate scores assigned by teams

Student	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8	Team 9	Team 10
Ania	7	6	6	7	6	7	9	7	7	6
Dominik	6	9	9	7	9	7	8	7	7	4
Łukasz	14	14	14	15	16	12	17	16	16	12

When the 'shaded' scores are removed from the dataset, the range of scores assigned by teams is reduced from 3–5–5 to 1–2–3 (for Ania, Dominik and Łukasz, respectively), which represents a considerable improvement. While no such doubts could be claimed to pester the 1-point discrepancy, I am not entirely sure, however, whether the 3-point difference could still be considered 'minimal'. Considering the stakes involved in the exam, I would argue to the contrary; after all, 3 points corresponds to a 15% difference in the score printed on the *matura* certificate. Yet in the face of the limited dataset used in the study, gauging the value of discussion as a score resolution method *vis-à-vis* the requirement specified in point (2) above must be withheld until more data is available.

In order to investigate the efficacy of discussion as a score resolution method in the EXAM with respect to the requirement specified in point (3) above I compared the aggregate scores arrived at by raters in the process of discussion with alternative aggregate scores obtained by averaging the aggregate scores assigned by individual raters prior to discussion (i.e., adopting the parity model). One obvious advantage of the parity model over the discussion model is that it speeds up the process of test administration by excluding from it the time needed to review student responses. On the other hand, a major disadvantage of this model is that it focuses on the aggregate scores only, which have already been shown to be arrived at using widely varied paths. However, considering the fact that it is only the

aggregate score that is reported to the examinee and the general public, adopting the parity model for the EXAM would appear to be a viable alternative to discussion.

Table 3 presents the *aggregate scores* assigned to the three students (Ania, Dominik and Łukasz) in result of discussion in a team and by averaging the original scores awarded by the members of each respective team. Table 4 shows the basic descriptive statistics computed for both types of aggregate scores. Note that the average scores were rounded to the nearest integers according to the following rule: .33 was rounded down, .66 was rounded up.

Table 3: Aggregate scores assigned in a process of discussion and by averaging the original scores of team members

Student	Score	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8	Team 9	Team 10
Ania	Disc.	7	6	6	7	6	7	9	7	7	6
	Aver.	**8**	**7**	**7**	7	**7**	**8**	8	7	7	**7**
Dominik	Disc.	6	9	9	7	9	7	8	7	7	4
	Aver.	**8**	8	9	**8**	8	7	8	7	**8**	**6**
Łukasz	Disc.	14	14	14	15	16	12	17	16	16	12
	Aver.	14	13	14	14	15	**14**	16	15	15	**13**

Note. 'Disc.' refers to scores assigned in the process of discussion. 'Aver.' refers to scores obtained by averaging the original scores of individual team members.

Table 4: Descriptive statistics for ratings assigned by teams in a process of discussion and by averaging the original scores of team members

Statistics	Discussion ratings						Averaged ratings					
Student	*n*	*M*	*SD*	*Min*	*Max*	*Range*	*n*	*M*	*SD*	*Min*	*Max*	*Range*
Ania	10	6.80	.92	6	9	3	10	7.40	.69	7	9	2
Dominik	10	7.30	1.56	4	9	5	10	7.70	.82	6	9	3
Łukasz	10	14.60	1.71	12	17	5	10	14.30	.95	13	16	3

As can be seen from Table 4, even without eliminating any 'extreme' ratings (cf. Table 2), the descriptive statistics for the aggregate scores obtained using the parity method are much better for the three performances in the study than those for the final scores arrived at in the process of discussion. Both the range and standard deviation values were reduced, indicating that the team ratings exhibited more homogeneity. Furthermore, in 12 instances, using the parity model resulted in a higher final score than when discussion was used and, most importantly, the scores previously identified as 'extreme' (cf. Table 2) were 'eliminated', thanks to which the respective team's 'new' (averaged) ratings better fitted with the 'new' ratings of other teams. It would thus appear that, at least for the dataset in this study, the parity model functions better and exhibits more of the desired characteristics than the discussion model. Considering the importance of psychometric properties of scores in high-stakes assessment contexts, this finding somewhat undermines the value of discussion as a score resolution method. The provisos addressed above still hold, however.

Finally, when the two types of final scores, i.e., formed in a process of discussion and by averaging, were correlated with the expert-criterion scores, it turned out that the correlation coefficients were similarly high for both types. Both Pearson's r and Spearman's ρ coefficients were computed, considering the small sample size their value should not be overestimated, however. For the discussion scores, the coefficients were: $r_{Dis:Exp}$=.932, p<.000; $\rho_{Dis:Exp}$=.795, p<.000; for the averaged final scores, they were: $r_{Ave:Exp}$=.968, p<.000; $\rho_{Ave:Exp}$=.796, p<.000. The fact that the correlation coefficients between the two types of operational ratings and expert-criterion scores were similar would appear to suggest that the parity model is in no way inferior to the discussion model. Moreover, in the light of the obtained descriptive statistics, as well as in view of the fact that the parity model takes into consideration the individual ratings (thus individual rater perspectives) in equal measures, this

model could even be claimed to exhibit more desirable characteristics than the discussion model.

All in all, although it was observed to iron out a number of discrepancies in the scores assigned by individual raters, discussion cannot really be claimed to function very effectively as a score resolution method in the EXAM, at least as far as I can tell on the basis of the limited dataset which I worked on. Discussion proved largely unsuccessful in yielding scores which would be identical or only minimally different across teams. The results of this study would appear to suggest that the parity model might work better as a score resolution method in the EXAM, although more detailed studies would need to be conducted to provide an answer to this issue.

Conclusion

One of the problems inherent in double marking of essays and spoken productions appears when the two (or more) discrepant scores must be somehow reconciled prior to reporting the operational score to the public. The method of arriving at this final score becomes a particularly vulnerable one in the case of high-stakes tests whose results are likely to determine examinees' fates.

The results of the small-scale study documented in this paper appear to indicate that in the case of the specific exam providing the context for the research (with all its idiosyncratic characteristics), discussion does not function satisfactorily as a score resolution method. Considering the fact that the primary interest of examinees is that the assessment is equitable and that their scores would be the same no matter which team scored their performance, the teams were not found to exhibit the desired characteristics; they were simply not interchangeable. Although some discrepancies between scores assigned by individual raters for individual items/sub-tasks/criteria *were* smoothed out in the process of discussion, the operational scores still varied substantially across the teams. Furthermore, qualitative analyses of the TD transcripts showed that the problems experienced by individual raters were all but alien to teams, whose ratings were found to vary for the same reasons which we identified as troublesome in the case of individual raters. Discussion in teams did not eliminate discrepancies in scores stemming from a lack of clear procedures for dealing with largely predictable situations (e.g., interlocutor support); it did not alleviate the consequences resulting from the misunderstanding of the rating scale descriptors; it did not safeguard against scores being awarded in result of heeding a whole array of criterion-irrelevant information.

One necessary proviso that needs to be made at this point is that the above conclusion only holds with regard to the EXAM 'as it stands' (the EXAM being scored by both trained and untrained teachers, lack of proper rater training and moderation sessions, poorly developed rating procedures, etc.). The results only bear to show that probably no score resolution method will be successful in and of itself in eradicating the variability that exists at the level of individual raters. It cannot be ruled out, however, that if proper training were provided and the raters were given guidance in the goals of discussion and sufficient practice in its application, discussion may have proved to act more than satisfactorily. As it stands, averaging the three individual scores appears to have worked better given the operationalisation of 'better' specified in response to research question (3). The practical aspect of adopting the parity model over discussion (less time necessary for obtaining the operational score, cost reduction) must not be ignored either.

Of course it could be claimed that the findings obtained with regard to teams are hardly generalizable owing to the small sample size used in the study, and I am aware of this limitation, as I am of many others (e.g., all participants were highly motivated volunteers).

Yet the results of the analyses may be – quite safely, I believe – treated as signposts, indicating further research avenues. That further research is desperately needed with respect to teamwork is obvious. Arguably, the observed lack of interchangeability amongst the teams is one of the most disheartening findings obtained in my study; disheartening, because it directly concerns the *operational* score, which is the score on which so much depends: passing or failing the EXAM, being accepted into university or not. The lack of interchangeability amongst teams leads to a situation in which the operational scores must not be interpreted in any but the most general manner, and their use had better be ceased altogether.

References

Clauser, B.E., Clyman, S.G., & Swanson, D.B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement*, 36, 29-45.

Davidson, F., & Lynch, B.K. (2002). *Testcraft. A teacher's guide to writing and using language specifications.* New Haven and London: Yale University Press.

Dubiecka, A., Szaleniec, H., & Węziak, D. (2006). Efekt egzaminatora w egzaminach zewnętrznych. In B. Niemierko & M.K. Szmigel (Eds.), *O wyższą jakość egzaminów szkolnych* (pp. 98-115). Kraków: Grupa Tomami.

Eckes, T., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., & Tsagari, C. (2005). Progress and problems in reforming public language examinations in Europe: cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France and Germany. *Language Testing*, 22(3), 355-378.

Johnson, R.L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, 13(2), 121-138.

Johnson, R.L., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication*, 18(2), 229-249.

Johnson, R.L., Penny, J., Fisher, S., & Kuhs, T. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education*, 16(4), 299-322.

Johnson, R.L., Penny, J., Gordon, B., Shumate, S.R., & Fisher, S.P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly*, 2(2), 117-146.

Linacre, J.M. (2005). *FACETS Rasch measurement program.* Chicago: Winsteps.com.

Linacre, J.M. (2006). *A user's guide to FACETS Rasch-model computer programs.* Chicago: Winsteps.com.

Myford, C.M., & Wolfe, E.W. (2002). When raters disagree, then what: examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement*, 3(3), 300-324.

Smolik, M. (2007). *Investigating scoring validity. A study of the 'nowa matura' speaking exam in English at the basic level.* Unpublished doctoral dissertation, Maria Sklodowska-Curie University, Lublin, Poland.

Weir, C.J. (2005). *Language testing and validation. An evidence-based approach.* Basingstoke: Palgrave Macmillan.