0204

Group dynamics in determining 'gold standard' marks for seeding items and subsequent marker agreement.

Beth Black, Milja Curcin
Cambridge Assessment, Cambridge, United Kingdom

Background

Online marking has necessarily entailed new possibilities and practices for monitoring marking. Along with other UK awarding bodies, OCR uses a system of seeding candidate responses into each examiner's marking allocation against which their marking accuracy is judged. Each seeding item is therefore marked by multiple examiners. The 'gold standard' or 'definitive' mark for each seeding item is determined by a team of senior examiners, led by the Principal Examiner, at a face-to-face meeting called the Standardisation Set Up meeting (SSU).

Since the decisions about definitive marks are made by a group of examiners, these decisions may be affected by 'group dynamics'.

Research Questions

We were particularly interested in the processes involved in determining the definitive mark for the seeding items and whether or not it would be possible to predict subsequent marker agreement (with the definitive mark) from an observation of the meetings and coding of specific meeting features. We also wished to discover whether there was any evidence that these decisions were subject to 'groupthink'.

Methods

Much research has investigated group dynamics and their impact upon the quality of group decisions. Many studies on conformity have found that individuals change their opinions when they find out what the majority opinion is in their group (e.g. Asch, 1951; Deutsch and Gerard, 1955). One possible consequence of conformity is group polarisation (Moscovici and Zavalloni, 1969), which refers to members of a group adopting a more extreme position. Groupthink is an extreme example of group polarisation (Janis, 1972) whereby independent thinking is lost to group cohesiveness and can bring about irrational decisions. Work on groupthink has often suggested that there are quite stringent antecedents for groupthink to emerge, though recent work has suggested that groupthink is far more ubiquitous than originally conceived (Baron, 2005). Correspondingly, there is also evidence that the presence of a dissenting minority can improve the quality of group decisions through greater consideration of alternatives and integration of multiple perspectives (e.g. Moscovici, 1976).

However, there is also some evidence that more cohesive groups tend to be more productive (Kerr and Tindale, 2004). Thus, this research indicates potentially contradictory outcomes for SSU whereby discussions with high contention and those with low contention might, for differing reasons, both be associated with 'good' decisions.

The SSU meeting could also be viewed in terms of leadership styles (e.g. democratic versus autocratic, Gastil, 1997). The literature regarding the impact of different leadership styles on group productivity is somewhat equivocal.

In this research, two observers attended five SSU meetings (3 GCSE and 2 AS units of a diverse range of subjects) in which the definitive marks for 2,025 seeding items were determined. For each seeding item these observers recorded:

1. Discussion time (seconds).
2. Contention level, (5 point scale) - the degree to which there was difficulty in agreeing upon a definitive mark due to differences in opinion between different examiners.
3. Democracy level, (5 point scale) - the degree to which the views of different panel members were encouraged, allowed and discussed.

We collected data on subsequent mark agreement for each seeding item to ascertain whether there was any relationship with the above meeting features.

Frame

The analysis was based upon a dataset of 2,025 seeding items, 168 markers and a total of 92,071 marking events.

The relationship between meeting features and subsequent exact marker agreement (P_0, see Bramley, 2007) was explored through a series of correlation analyses and analyses of interaction (e.g. ANOVA), taking into account relevant information about each of the seeding items such as question features (e.g. item type - objective, short-answer etc.), mark scheme features (e.g. mark scheme approach - objective, levels-based or points-based) and response features (e.g. legibility).

In order to explore whether there was any evidence for groupthink, we identified those items where the definitive marks determined at the SSU meeting were incongruent with the consensual (modal) marks of the subsequent markers.

Research findings

Overall, very high levels of exact marker agreement were found.

There were strong negative correlations between both discussion time and contention levels and subsequent marker agreement, indicating that longer discussion times and higher levels of contention were associated with lower levels of marker agreement. Indeed, of all the question, mark scheme and response features coded for, these were two of the strongest predictors of subsequent marker agreement.

It seems that the relationship between these two features and marker agreement arises because they are an expression, or composite, of many of the other features of the items, mark schemes and responses. Discussion and contention (which themselves correlate highly) tend to intertwine problems with peculiarities of the response, operationalisation of the mark scheme, the nature of the item (e.g. objective versus extended response) and so on. Discussion time and contention increase as a function of the difficulty of the decision making process (e.g. the number of competing rationales for awarding a different mark): subsequently, markers will encounter the same difficulties in their decision-making process for these items, but, in awarding a specific mark, may (legitimately) decide to resolve these issues differently. Thus the contention level is a product of the cognitive demands of the marking task, rather than high contention levels necessarily adversely affecting the soundness of the decision.

The third meeting feature observed, democracy, had less straightforward findings. The overall correlation between democracy rating and P_0 was non-significant (Rho = 0.019), though it was significant for two of the five examinations. This possibly suggests that there are conditions under which democracy may or may not be a good predictor of marker agreement and thus warrants further investigation.

For only a very small number of items (4%) was the definitive mark incongruent with the modal mark. In the majority of these cases, the discrepancy could be explained by differing but legitimate interpretations of the response and its match with the mark scheme. Thus there was no evidence of groupthink in SSU meetings.