A STUDY OF DISCREPANCIES IN THE ASSESSMENT OF PROBABILISTIC TASKS: WHY MIGHT TEACHERS GRADE AND EVALUATE INCONSISTENTLY A GIVEN ANSWER?

NABBOUT, Marie
Institut Libanais d'Educateurs -FSédu -Université Saint Joseph
Lebanon

In this paper we present part of a study carried out to identify Lebanese teachers' representations of probability as well as their teaching practices. We compare grades and judgments that teachers attribute to fictitious students. Our result show inconsistencies in grading on the same teacher: discrepancy between the quantitative judgment (grade) and the qualitative judgment that he attributes to the same answer. The comparison between the quantitative judgments (grades) and the qualitative judgments reveals a great diversity among teachers: a convergence in grading can hide very different qualitative judgments. This comparison contributed largely to the study of certain representations and practices of teachers, in particular those concerning the concept of independence of events.

INTRODUCTION

Discrepancies in grading amongst evaluators is a fact, the aim of this paper is to go beyond this fact and study the underlying reasons to these discrepancies when they exist and to identify teachers' representations which might be the original source. Probability problems are good examples to study such discrepancies between mathematics teachers, and that to the unpredictable nature of probability, the role of intuitions as well as the fragility of answers (Maury & Nabbout, 2003).

Our analysis is based on the *reference model* (modèle de référence) (Noizet & Caverni, 1978), and the *theory of elaboration of conflict* (théorie de l'élaboration du conflit) (Perez & Mugny, 1993).

THEORETICAL BACKGROUND

Reference model

Noizet, Caverni & al. (1969, 1978) affirm that discrepancies in grading are not only subject to the personality of the assessor but to the task itself and its variables. The authors describe the behavior of assessors as follows: the assessor extracts from the given answer a number of elements that he compares to his own *reference model* embedded in his cognitive structures (Noizet & Caverni, 1978). Accordingly, the assessment task is transformed from being an absolute estimation to a comparative estimation. The *reference model* is formed prior to the assessment act, but it is subject to constant modifications. It is made of a "norm", an expected "product" and a position on a measuring scale. The general hypothesis of Noizet & Caverni is that two assessors not having the same "norms" or expected "product" would refer to different models of reference and thus assess differently a same given answer. Several factors were identified by the authors as influencing the reference model, particularly prior information about the students, the order of copies (assessors tend to over assess the first copies and under assess last copies); as well as the experience of assessors (novice teachers are more indulgent then veterans). Consequently, grades are never absolute, but are relative to a reference which is subjectively present and modifiable during the assessment. The feedback that the assessor gets after correcting a copy provokes a change in his reference model and accordingly changes the correction of the following answers. In view of that, an answer might be over (or under) assessed depending on the previous corrected copy.

The theory of elaboration of conflict

The Theory of Elaboration of Conflict (TEC) postulates that social influence (modification of a behavior resulting from the exposure to an influential source) results from a divergence of opinions with others (*cf.* Pérez & Mugny, 1993). An uncertain subject, who finds himself in a situation where his personal aptitudes and competencies are in question, is ready to

refer to any information that might help him out. Uncertain subjects are those who are unable to justify or verify their judgments in a given situation. The TEC distinguish between four types of tasks, according to the nature of the task and the influential source, we are only concerned by the TAP (Task of APtitude) which reveals the personal competencies of individuals. Authors distinguish three cases of TAP:

- the individual knows that there exists an answer that he ignores, and it is impossible for him to find it;
- the individual possess two (or more) incompatible answers that seems correct to him; he doesn't know which one to choose;
- the individual doesn't know the answer and try to finds it.

In a TAP, individuals are concerned by correcting any wrong image of self, and any divergence between their answers and the source doubles uncertainty. Facing a minor influential source (students in our case), the fear becomes double: a teacher cannot accept a wrong answer because it jeopardizes his skills, and cannot reject an answer without being sure that it is erroneous. Conflicts are resolved, according to TEC, depending on the source. Facing a high influential source (norms in our case), conflicts are resolved by imitation; facing a minor influential source (students in our case), conflicts are resolved by examining all details in the answer.

GENERAL FRAMEWORK

The data on which our work is based are collected from a research (Nabbout, 2006) that aims at studying some representations of Lebanese French speaking high school mathematics teachers in probability, and their methods of teaching. The research consists of four individual interviews with sixteen Lebanese mathematics teachers. Data studied in this paper were collected from the third interview where teachers had to evaluate students' work. We studied, in simulated situations, the activities of those teachers in different tasks related to assessment: production of tests, grading and evaluating "fictitious" students' answers. Teachers were asked to correct and grade students' answers on eight different problems and where, for each problem, four answers were proposed. Situations were presented one by one to teachers who were asked, after grading, to explain their work and to evaluate (give a qualitative judgment) the corrected copies.

In this paper, we studied discrepancies in the assessment of teachers (within and between teachers) and we try to relate it to their representations of probability and to assessment. We analyzed and compared between their assessment and their evaluation of students' answers, particularly for problematic cases where answers cannot be judged right away.

THE TASKS

Each of the sixteen teachers was asked to correct eight different problems, four copies for each problem. The thirty two copies are designed by SiEj, where "i" designates the number of the problem and "j" that of the answer (e.g. S4E3 stands for the third answer given to problem 4). Problems were about simple probability (S2, S8); independent evens (S1, S4, S5, S6, S7) and conditional probability (S3).

Concerning independent events, it is important to distinguish between:

- "*intuitively independent events*": two events are said to be intuitively independent when "they are not influenced one by the other", that is when they are associated to experiences occurring successively and where the independency is postulated (*cf.* Steinbring, 1985). Such events are as well called "chronologically independent events" (*cf.* Maury 1985) when the situation consists of successive "independent" events in the naïve sense of the term) (S1, S5, S7)
- and "*stochastically independent events":* two events A and B are said to be stochastically independent when "$P(A \cap B) = P(A) \times P(B)$". In this case, no reference is made to any experiments or to any chronology, and the independency is only defined by the mathematical formula) (S4, S6) (Maury, 1986; Steinbring, 1985; Nabbout, 2005; 2006).

The independence of events is the core element of the study and in view of that additional importance was given.

Answers were classified on three categories: correct, wrong and other (cf. Maury & Nabbout, 2003).

- The category "*correct*" covers intuitively correct answers and formally correct answers (answers which correspond to the official requirements set by the program).
- The category "*wrong***"** covers erroneous answers where one can identify the origin of the mistake; (e.g. confusion between incompatibility and independence).
- The category "*other*" covers cases where a correct element is jointly present with a wrong or out of context argument.

In a given answer, we distinguish between *procedure* and *result*. We talk about *conform result*, the cases where the *result* is correct, i.e. conform to the expected mathematically correct answer.

RESULTS

Grades

Statistical study of grades shows that problematic cases (where major discrepancies among grades were observed) occurred with correct intuitive answers, wrong answers with conform result, and with "other" cases particularly with non-conform results. Convergence of grades (range < 2) occurred with wrong answers with non-conform result. Neutral cases (2< range<3) occurred mostly with correct formal answers. The outlier values are due to the teachers' profile or to his reference model.

The conformity of the answer to the result seems to be a disruptive factor to teachers, since such cases were divergence grades.

Evaluative judgment

In order to study the qualitative evaluation of teachers and to compare it with grades, we classified their judgments in a three category nominal scale (wrong, average, and correct). We split these categories into seven subcategories and we attributed codes going from "completely wrong" (0) to "completely correct" (7). Similar discrepancies were observed on the evaluative judgment.

Comparison between grades and evaluative judgment

Results show that teachers *may evaluate a given answer the same way, but would disagree over the grade. As well, they may grade a given answer the same way but would evaluate it differently*. The main factor for these discrepancies between grades and judgments is the conformity of the result and the intuitive non standard answers. *The conformity of the result to the expected answer is a main disruptive element* for teachers. A consensus is observed with wrong answers having non-conform results, this consensus fade away to wrong answers having conform result. Agreement in judgments is observed with formal correct answers and with wrong answers; the difficulty is with partially correct (or wrong) answers.

We think that teachers tend to refer to formal norms while grading in order to be on the safe side; but seem to be freer while judging and express their point of view even if it contradicts the attributed grade; which explain a high qualitative judgment to a low grade.

Teacher's assessing profile

Some discrepancies are due to the *assessing profile of teachers*, a profile that we elaborated for each teacher taking into consideration all the grades used in the correction of the thirty two copies. For some teachers, it is hard to find a correct answer, they never attribute a full grade, and others avoid average grades and rather give full or null grades (cf. P13 - figure 1).

Grades and judgment were significantly correlated for some teachers (cf. P8 - figure 2), but not correlated for others. For some teachers, evaluation would go above the grade (cf. P16 - figure 3), it is the other way round for others and not significant for the rest of them (cf. P5 - figure 4).

Another result worth mentioning, teachers tend to be strict (respectively indulgent) with grades having an answer evaluated as "good" (respectively bad).

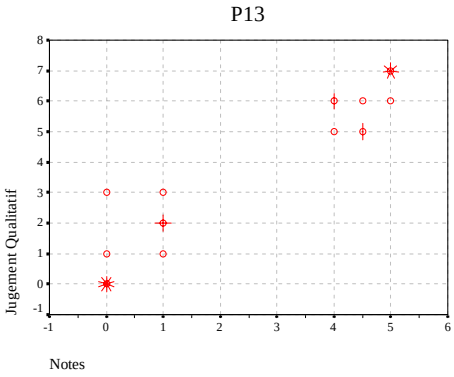

Figure 1. Teacher avoiding average grades

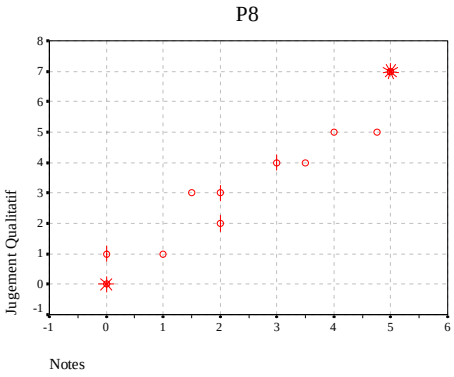

Figure 2. Correlated grades and judgment

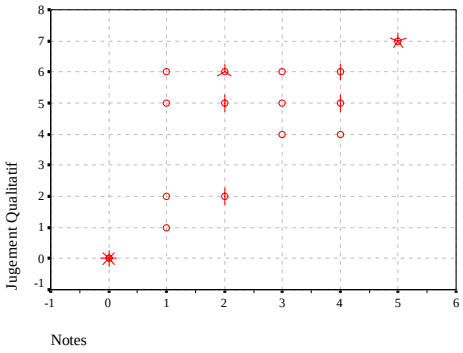

Figure 3. Judgments exceeding grades

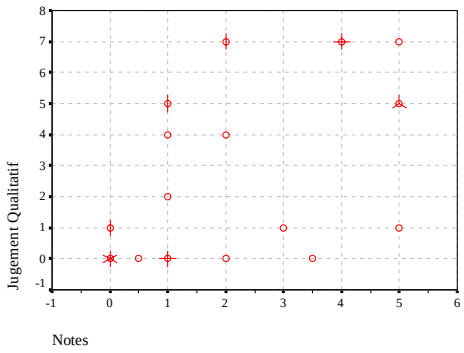

Figure 4. Not correlated grades and judgment

Construction of the reference model: example

Teachers in our study never preset a scale; their rare reference to a scale occurred as a pretext to justify their grades. Solving the problem preliminary to the correction was an automated action for some teachers, for others it occurred with doubtful cases. The study of the construction of the reference model by teachers was the way helping us understand discrepancies amongst grades and judgment of the same teacher; and allowing us to identify elements of his representation of the underlying mathematical concept. We illustrate it by the following case study of teacher P1. P1 is a teacher on whom the *implicit confusion* between incompatibility and independence was identified. Implicit confusion between incompatibility and independence occurs when teachers try to make sense of the independency of events, they tend to use paraphrases of the definition, and they considered that when there is no influence of one event on the other, then the two events are independent. This "no influence" is later replaced by "no common element" (cf. Nabbout, 2005; 2006).The implicit confusion between incompatibility and independence (erroneous representation for independent events) may coexist with identification of explicit confusion between incompatibility and independence on students.

- Prior to his correction to the problem S4, P1 solved the problem. He seems to be conscious of the existence of cases where the independence of events is not intuitive and where he must refer to the calculation; or might it be to verify his intuitive answer (which hides a misconception). At this stage, we can tell that *his reference model is the calculation.* We think that P1 doubted the correctness of these answers, but the conformity with his (erroneous) calculation lead him to accept them.

S4: *A fair die is tossed once. Consider the following events, A: "the number is even", B: "the number is greater or equal to 4", and C: "the number is a multiple of 3".*
1) Are A and B independent? Justify your answer.
2) Are A and C independent? Justify your answer.

- P1 did an erroneous calculation, according to which S4E1 is a conform answer. He then accepts S4E1 and S4E2 (confusion between incompatibility and independence), two "intuitive" yet wrong answers, which are conform to his *wrong* calculation. At this stage, the *reference model becomes the calculation or S4E1.*
- P1 refuses S4E3 (the events are not independent because they have common elements) which contradicts his new reference model.
- While correcting S4E4 (based on calculation), P1 realizes his calculation mistake. He repeats his calculation and accepts the given answer. The model of reference is again the calculation.
- After correcting problem S6 (similar to S4), P1 admits that "the existence of a common element between two events is not a reason for non independency of these events". He re-corrects S4 and refuses S4E1 and S4E2 without being able to explain the reason.

The TEC explains well what happened with this teacher at this level: correcting is a TAP, reflecting teachers' competencies in the subject matter, P1 accepted S4E1 believing its conformity with the calculation (the norm). Once he discovers the discrepancy between S4E1 and the calculation, he rejects S4E1 and refers only to the norm.

CONCLUSION

Our results show that the analysis of students' answers is not a mere professional action accompanying a grading task, but a reaction attributed to the answer. When teachers are able to decide answers (correct or wrong) they never try to analyze it; teachers tend to analyze students' answers in case of difficulty. We interpret this action by referring to the TEC (Theory of Elaboration of Conflict): teachers resolve conflicts emerged by minor influential source (students) by examining all details in the answer since their acceptance or rejection of answers reflect their skills which they cannot jeopardize.

Results show that the grade does not reflect the qualitative judgment of teachers. An important discordance was observed between grades and judgments given to the same answer between teachers and within the work of the same teacher: a convergence on the same grade by different teachers may hide discordant judgments for the same copy; as well, discordances are observed between grades and evaluations given all the way, by the same teacher. Teachers have a tendency to refer to the norms while grading, but they feel more "free" to express themselves while evaluating the copy, which explains a high evaluation with a low grade. Teachers' attitude towards correct intuitive answers explains divergence between grades and evaluations; intuitive answers are mostly accepted in typical intuitive situations of probability (e.g. draw without replacement).

The conformity of results seems to be a disconcerting factor to teachers, obliging them to analyze deeply students' answer. In problematic cases, teachers have the tendency to base their judgment on the conformity of the answer.

REFERENCES

Maury, S. (1985). Influence of the question in a task on the concept of independence. *Educational Studies in Mathematics*, 16, (pp. 283 – 301).

Maury, S. (1986). Contribution à l'étude didactique de quelques notions de probabilité et de combinatoire à travers la résolution de problèmes. *Thèse de doctorat d'état – Montpellier*.

Maury, S. & Nabbout, M. (2003). Analyse de quelques données concernant l'évaluation de travaux d'élèves en probabilités par cinq professeurs de mathématiques francophones enseignant en lycée. *Colloque EMF (Espace mathématique Francophone)*. Tunisie.

Nabbout, M. & Maury, S. (2005). Teachers' representations of independent events: What might an attempt to make sense hide? *CERME IV Proceedings, Spain.*

Nabbout, M. (2006). Enseignement des probabilités en classe terminale au Liban: étude de représentations et de pratiques de professeurs dans des situations aménagées. *Thèse de doctorat – Université René Descartes – Paris5 – France.*

Noizet, G. & Caverni, J.P. (1969). Pour une docimologie expérimentale. *Bulletin de Psychologie 22, n° 276*, (782-787).

Noizet, G. & Caverni, J.P. (1978). Psychologie de l'évaluation scolaire. *Presses Universitaires de France*.

Pérez, J.A. & Mugny, G. (1993). Influences Sociales : la théorie de l'élaboration du conflit. *Delachaux et Niestlé.*

Steinbring, H. (1986). L'indépendance stochastique : un exemple de renversement du contenu intuitif d'un concept et de sa définition mathématique formelle. *Recherches en Didactiques des mathématiques, 7, n°3,* (pp.5-50).