Supplementary material for the paper Classifying Students' Meta-Cognitive Comments

A Classification Scheme

Table A.1 summarizes the 9 consolidated categories at the output of the classification scheme generation process, their definitions and typical keywords. The first three categories concern cognitive evaluations of the test results. The fol-

Category	**Definition**	**Typical Keywords**
MonCal.Facts.RF	Factual evaluation of the test results, in terms of number/quantity of correct/incorrect answers	Correct, Incorrect, Right, False
MonCal.Facts.DC	Factual evaluation of the test results, in terms of number/quantity of uncertain knowledge, well-founded knowledge, unexpected errors, presumed errors	Uncertain knowledge, Well-founded knowledge, Unexpected errors, Presumed errors*
MonCal.Interpret	Interpretations of the test results beyond quantitative evaluation, reflecting students' subjective interpretations of their performance	Prudence, Imprudence, Doubt*
BDS.Emotions	Comments expressing emotions felt by the student during and after the test and their implications	Stress, Anxiety, Fear, Disappointment, Satisfaction
MotOrient.Value	Comments about interest in test type (true false questions with degres of certainty) and their perceived value to learners	Interest, Utility, Allow, Useful
MotOrient.SelfEff	Comments on students' self-efficacy and confidence related to the tested domain	Confidence, Abilities, Overestimation, Underestimation, Realism, Competency
DomainKldg	Comments identifying missing or achieved disciplinary concepts/domains and their degree of acquisition	Gaps, Concepts, Lesson, Subject, Chapter, Physics
StratKldg	Comments about study tactics, learning strategies and theirs impacts	Forgetting, Misreading, Difficulties, Attention, Inattention, Lack of time
CTRL	Comments concerning forward planning ; projection into future tasks	Next time, Next Test, Future, have to, should

Table A.1. Classification Categories and Keywords.
*These keywords reflect the specific vocabulary used in our context to provide post-test feedback: Uncertain knowledge (correct answer and low certainty), Well-founded knowledge (correct answer and high certainty), Unexpected errors (incorrect answer and high certainty), Presumed errors (incorrect answer and low certainty)

lowing five categories deal with students' comments showing that the test had an impact on their cognitive conditions: emotional dispositions, motivational factors, self-efficacy, awareness of gaps or changes in domain knowledge, enhanced knowledge of their study tactics and learning strategies. The final category concerns forward planning, i.e. students' reflections on the attitudes and behaviors to adopt in future learning tasks. These categories represent different metacognitive processes in the COPES model (Winnie and Hadwin, 1998).

B Classification Results for Corpus 2

B.1 Agreement between two human coders

Fig. A1 illustrates the number of agreements and disagreements between the two human coders involved in the classification of Corpus 2. The numbers in the diagonal (green cells) are the number of agreements, off-diagonal numbers indicate disagreements. The Inter-Coder-Reliability measured by Cohen's Kappa is 0.67. The high number of disagreements (10) between StratKldg and DomKldg can be explained by the fact that one coder assigned comments like "*I realized that I still hadn't revised many of the topics covered by the test*" to the DomKldg category, while the other classified it in the StratKldg category. This type of comment, while not explicitly identifying a specific notion to be learned, implicitly suggests it, but it can also be interpreted as an unsuitable pre-test learning strategy. This is a typical example of how the ambiguity of some comments makes it difficult to establish a ground truth. Other examples concern the MonCal.Interpret and MotOrient.SelfEff categories, as in the comment: "*I have too much confidence in incorrect answers*", which can be classified both as the student's interpretation of the test result received and as an adjustment of perceived competence (self-efficacy).

The human coders agreed to choose a category for contentious cases in order to establish a ground truth. In many cases, the other choice was also acceptable. This suggests an avenue to explore for another approach to classification, which could be to allow the same comment to be classified in two different categories. The disadvantage of this approach is that it would create repetition in the comment classification table and complicate analysis.

	MonCal.Interpret	MonCal.Facts.RF	MonCal.Facts.DC	BDS.Emotions	MotOrient.Value	MotOrient.SelfEff	DomKldg	StratKldg	CTRL	SumRow
MonCal.Interpret	18	0	4	0	0	7	1	1	0	31
MonCal.Facts.RF	0	26	0	0	0	0	1	0	0	27
MonCal.Facts.DC	1	0	47	0	0	0	0	1	0	49
BDS.Emotions	2	0	0	3	0	0	0	0	0	5
MotOrient.Value	0	0	0	0	5	2	1	2	0	10
MotOrient.SelfEff	0	0	0	0	0	13	0	0	0	13
DomKldg	0	0	0	0	0	0	30	0	3	33
StratKldg	0	0	0	0	0	0	10	27	2	39
CTRL	1	0	0	0	0	0	1	6	20	28
SumCol	22	26	51	3	5	22	44	37	25	

Fig. A1. Inter-Coder-Agreement Matrix between the two human coders.

B.2 Agreement between LM Phi-4 and the Human reference classification

Fig.A2 illustrates the alignment between the final human classification (after coder consensus) and the classification achieved with the LM Phi-4. The rows correspond to the human reference and the columns to the LM classification. The Inter-Coder-Reliability measured by Cohen's Kappa is 0.77 and the accuracy 0.8. Regarding disagreements between LM Phi-4 and human classification, we find discrepancies similar to those between human coders. Comparable difficulties in deciding which category is most appropriate appear between StratKldg and CTRL, and between DomKldg and StratKldg. A total of 47 comments were classified differently by the LM than in the reference. We evaluated each of them to judge whether the alternative classification was acceptable (as in the examples given above for human disagreements). This was the case for 29 comments, so only 18 were actually misclassified. If we recalculate the accuracy on this basis, it rises to 0.92.

	MonCal.Interpret	MonCal.Facts.RF	MonCal.Facts.DC	BDS.Emotions	MotOrient.Value	MotOrient.SelfEff	DomKldg	StratKldg	CTRL	SumRow
MonCal.Interpret	17	0	3	0	0	1	0	0	2	23
MonCal.Facts.RF	0	26	0	0	0	0	0	0	0	26
MonCal.Facts.DC	2	1	48	0	0	0	0	0	0	51
BDS.Emotions	0	0	0	2	0	1	0	0	0	3
MotOrient.Value	1	0	0	0	4	2	0	0	1	8
MotOrient.SelfEff	3	0	0	0	0	18	0	0	0	21
DomKldg	0	1	0	0	0	2	27	2	2	34
StratKldg	1	1	0	0	0	0	9	25	10	46
CTRL	0	0	0	0	0	0	2	0	21	23
SumCol	24	29	51	2	4	24	38	27	36	

Fig. A2. Inter-Coder-Agreement Matrix between human (rows) and Phi-4 (columns).