Grading Without the Answer Key

Asal Kc

05/08/2026

Abstract

Consider a multiple-choice exam whose answer key has been lost, leaving only the students' answer sheets. We show that the sheets alone already determine a great deal about the scores. If the key marks option ℓ correct on a question, the students who earn that mark are exactly those who wrote ℓ, so the total number of marks awarded to the class is a sum of *group sizes*, one group chosen per question. From this we obtain an exact description of the set of score vectors any key could produce, a ceiling on the class average computable in a single pass, and the fact that any two students disagreeing on D questions contain one with at least $D/2$ wrong. We also prove a matching impossibility: between any two students who differ at all, every score gap consistent with parity is achievable by some key. The aggregate is heavily constrained while the ranking is not constrained at all. One can therefore prove from the sheets that the class did badly, and never prove who came first. Nothing beyond counting is used, and no assumption of any kind is made about how the students answered.

1 Setup

An exam has n questions. Question i offers a set of options $\mathcal{A}_i$ with $|\mathcal{A}_i| \geq 2$. Each of m students submits an answer sheet $f_j \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$, so f_{ji} is the answer student j gave to question i. An *answer key* is any y in the same set. The score of student j under key y is

$$s_j(y) \;=\; \big|\{i : f_{ji} = y_i\}\big|.$$

We are given $f_1, \ldots, f_m$ and never y. Everything below is a consequence of one observation. Question i sorts the students into groups according to what they wrote. For $\ell \in \mathcal{A}_i$ let

$$B_i(\ell) = \{j : f_{ji} = \ell\}, \qquad b_i = \max_{\ell \in \mathcal{A}_i} \big|B_i(\ell)\big|,$$

so the nonempty $B_i(\ell)$ partition the students, and b_i is the size of the largest group, basically being the number of students who gave the most popular answer to question i.

Proposition 1 (The grading identity)**.** *For every key y,*

$$\sum_{j=1}^{m} s_j(y) \;=\; \sum_{i=1}^{n} \big|B_i(y_i)\big|.$$

Proof. Count the pairs (j, i) with $f_{ji} = y_i$ in two ways. Grouping by j gives the left side. Grouping by i gives the right side, since the students scoring on question i are precisely those who wrote y_i. □

The key never appears in the identity except through which group it selects. Marking an exam is therefore, from the class's point of view, nothing but choosing one group per question and adding up their sizes.

2 What the sheets force

Theorem 2 (Class ceiling)**.** *For every key y, the class average score satisfies*

$$\frac{1}{mn}\sum_{j=1}^{m} s_j(y) \;\le\; \overline{C} \;:=\; \frac{1}{mn}\sum_{i=1}^{n} b_i,$$

with equality if and only if y_i is a most popular answer on every question. The quantity $\overline{C}$ is computed from the sheets in one pass.

Proof. Immediate from Proposition 1, since $|B_i(y_i)| \le b_i$ for each i and the choices on different questions are independent. □

If thirty students sit a paper on which no answer is ever given by more than twelve of them, then whatever the key says, the class average is at most 40%. Clearly, this is not just an estimate.

Corollary 3 (Two students)**.** *Suppose students j and k gave different answers on exactly D questions. Then their numbers of wrong answers sum to at least D. In particular, at least one of them has at least $D/2$ wrong.*

Proof. On each of the D disputed questions the key agrees with at most one of the two sheets, so at least one of the two students is wrong there. Summing over the D questions gives the claim. □

Corollary 4 (At most one strong student)**.** *Let $D_{\min}$ be the least number of disagreements between any two sheets. Then at most one student has fewer than $D_{\min}/2$ wrong answers.*

Proof. If two students each had fewer than $D_{\min}/2$ wrong, Corollary 3 applied to that pair would give a disagreement count below $D_{\min}$. □

Remark 5 (There is no matching floor). Proposition 1 also gives $\sum_j s_j(y) \ge \sum_i \min_{\ell\in\mathcal{A}_i} |B_i(\ell)|$, but this is usually 0 because if some question offers an option nobody chose, the key may select it and award nothing. A positive floor exists exactly when every option of every question was used by some student. Agreement among the students is never by itself evidence that they are right, so a question all m students answer identically contributes m to the ceiling and 0 to the floor.

3 Exactly which score vectors are possible

Theorem 2 constrains the sum of the scores. The sheets in fact determine the entire set of achievable score vectors, and the description is exact.

For each question i and option $\ell \in \mathcal{A}_i$, let $\chi_i(\ell) \in \{0,1\}^m$ be the indicator vector of the group $B_i(\ell)$, and put $V_i = \{\chi_i(\ell) : \ell \in \mathcal{A}_i\}$. Each V_i consists of the indicator vectors of the blocks of a partition of the students, together with the zero vector whenever some option of question i went unused.

Theorem 6 (Achievable scores)**.** *The set of score vectors realizable by some key is exactly the Minkowski sum*

$$S \;=\; V_1 \oplus V_2 \oplus \cdots \oplus V_n \;=\; \left\{\textstyle\sum_{i=1}^{n} v_i \;:\; v_i \in V_i\right\} \;\subseteq\; \mathbb{Z}^m.$$

Proof. Write $s(y) = (s_1(y), \ldots, s_m(y))$. Student j scores on question i precisely when $j \in B_i(y_i)$, so $s(y) = \sum_{i=1}^{n} \chi_i(y_i)$, which lies in S. Conversely each element of S is $\sum_i \chi_i(\ell_i)$ for some choice of options ℓ_i, and the key $y = (\ell_1, \ldots, \ell_n)$ realizes it. □

Every earlier statement is a corollary: Theorem 2 is the maximum of $\mathbf{1}^\top s$ over S, and Corollary 3 is the observation that on a disputed question no element of V_i has a 1 in both coordinates.

Theorem 6 is also an algorithm. The reachable sets $R_0 = \{\mathbf{0}\}$ and $R_i = R_{i-1} \oplus V_i$ are computed by dynamic programming over the questions, using at most $|\mathcal{A}_i|$ additions per stored vector, so deciding whether a claimed score vector is achievable is polynomial for any fixed number of students. Restricted to a pair of students it is a two-dimensional table and runs instantly.

Example 7. Four students, five questions, options $\{A, B, C\}$ throughout:

	Q_1	Q_2	Q_3	Q_4	Q_5
Student 1	A	A	A	C	A
Student 2	A	B	B	C	A
Student 3	A	A	C	C	B
Student 4	B	B	A	C	B
largest group b_i	3	2	2	4	2

Here $\sum_i b_i = 13$ out of $mn = 20$, so no key gives the class an average above 65%. Students 1 and 2 disagree on two questions, so at least one of them loses a mark. Question Q_4 illustrates the previous remark: unanimity contributes the largest possible amount to the ceiling while guaranteeing nothing at all.

4 What the sheets can never force

The constraints above are all aggregate. The following shows that nothing survives at the level of comparisons, which is the natural thing one would want to recover.

Theorem 8 (Rankings are unidentifiable)**.** *Let students j and k disagree on a set D of questions, $|D| = D$.*

(i) If $D = 0$ the two students score equally under every key.

(ii) If every question in D has $|\mathcal{A}_i| = 2$, then $s_j(y) - s_k(y)$ can be made equal to any integer t with $|t| \le D$ and $t \equiv D \pmod 2$, and no other value occurs.

(iii) If every question in D has $|\mathcal{A}_i| \ge 3$, then $s_j(y) - s_k(y)$ can be made equal to any integer t with $|t| \le D$.

In particular, whenever $D \ge 1$ there are keys under which j beats k and keys under which k beats j.

Proof. Questions outside D contribute equally to both scores and so cancel. On a question in D, the key contributes $+1$ to the difference if it equals f_{ji}, -1 if it equals f_{ki}, and 0 otherwise; the last is available exactly when a third option exists. The contributions on distinct questions are chosen independently, so the attainable differences are all sums of D terms drawn from $\{+1, -1\}$ in case (ii) and from $\{+1, 0, -1\}$ in case (iii), which are the stated sets. □

So the sheets place the class, but never the students. One may certify from the sheets alone that the average was poor, that a particular pair of students cannot both have done well, and that at most one student in a mutually disagreeing cohort was strong, and one can never certify that any student outscored any other.

5 Uses

Grading pipelines. The same check catches a misaligned key or a mis-scanned batch, which typically inflates or deflates scores in a way that leaves the ceiling behind.

Machine graders. Nothing above used any property of students. Replacing students by predictive models, questions by benchmark items, and the key by the ground truth, the results become assumption-free statements about model accuracy computable from released predictions.

6 Limitations

Theorem 2 bounds the average and hence the maximum, but not the minimum i.e. one excellent student among many poor ones is consistent with a low ceiling. Corollary 4 is the statement that constrains individuals, and it constrains all but one of them at once. Finally, the results are exact and worst-case by design and say nothing about which key is *likely*.