Calibration

Calibration is a second grade evaluation method, which compares student's marking abilities against that of the teacher's, using the provided example assessments.

Mechanism

Calibration is an evaluation sub-plugin. As of Moodle 2.2, the only evaluation sub-plugin is "best", or "comparison with best assessment" which compares markers against each other in order to evaluate the "best" grade.

Calibration is a more accurate method of assessment as it compares all markers against a single standard that is set by the teacher. These comparison assessments are the same as the existing example assessments in Workshop.

When the teacher selects the Calibrated method, they are given two controls to adjust the grading curve: "Comparison against example assessments" and "Consistency of assessment accuracy", known in short as Comparison and Consistency.

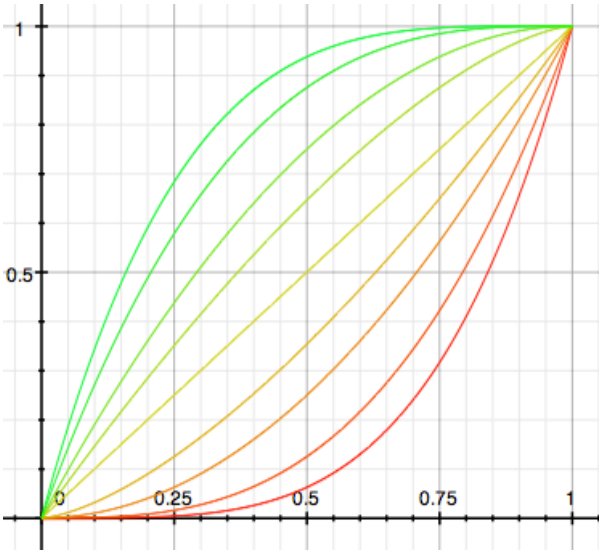

Fig. 1

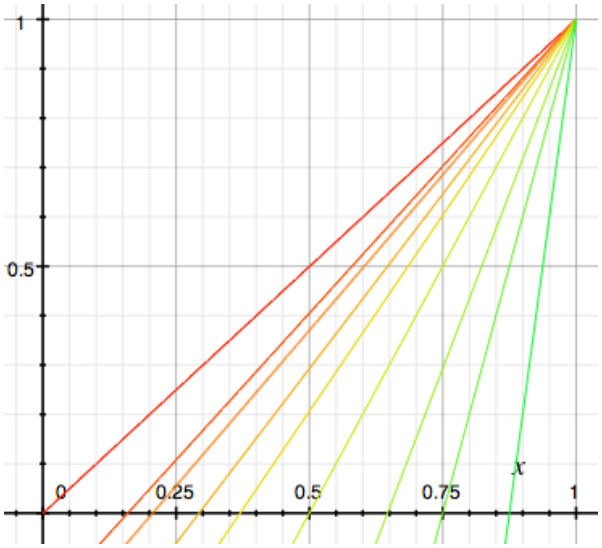

Fig. 2

These graphs represent the curves for comparison (fig. 1) and consistency (fig. 2). Red is stricter, green is more relaxed.

Comparison is straightforward: the x-axis is correlation, where 1 is an assessment completely identical to the example assessment and 0 is completely wrong; and the y-axis is the resultant score after scaling. For example, on the most relaxed scale, even an average 50% accuracy gives the marker a grade of around 95%, whereas on the strictest

setting, this gives them just over 5%, meaning they have almost no effect on the final marks of submissions the have assessed.

Consistency comes into play when there is more than one example assessment. It is a measure of how consistent a marker is. For example, if a marker scores 80% on two example assessments, they are more consistent than one who scores 100% on one and 50% on the other. The calculation is a little more complex; it follows the equation:

$$S' = (S \cdot (1 - AD)) \cdot C + S \cdot (1 - C)$$

where:

- S' is the resultant score
- S is the mean raw accuracy determined by the Comparison setting
- AD is the absolute deviation of the raw accuracy scores (remember there is more than one example assessment)
- C is the consistency factor set by the teacher, scaled according to the graph above

On the graph, the x-axis is the mean raw score, and the y-axis is the resultant calibration factor.

The setting here determines how much accuracy is a factor for highly accurate assessors. On the strictest setting, accuracy is a factor for everyone, although its import decreases as the mean raw accuracy decreases. On the default setting, an assessor must score over 37% for accuracy to even be a factor in their score, and on the most relaxed setting scores under 80% aren't penalized for inaccuracy at all.

Implementation

Calibration is implemented as an evaluation sub-plugin.

This first involved altering the evaluation form slightly as the evaluation method was a static field. The form is now derived from the selected evaluation plugin, which must have a field to select other plugins.

The plugin installs a single table, `workshopeval_calibrated`, which saves evaluation information state. It has four fields, `id`, `workshopid`, `comparison`, the most recently set comparison weighting on a scale from 1-9, and `consistency`, the most recently set consistency weighting on a scale from 1-9.

The function defined in the interface, `update_grading_grades`, is a fairly straightforward iteration of each of the markers within an assessment wherein each is given their calibrated grade which is then normalized to a score out of 100 and then the assessment records are updated with this score.

There are two internal tables, one of which relates to the curves given above, and the other is an internal scaling used for affecting how much the consistency factor affects scores, where one is strictest and nine is most relaxed.

	1	2	3	4	5	6	7	8	9
Curves	1/4	1/3	1/2	2/3	1	3/2	2	3	4
Dvn. Factor	2.0	1.75	1.3	1.0	0.75	0.5	0.35	0.28	0.2

The bulk of the work happens in `calculate_calibration_score`. This algorithm can be broken down as follows:

1. For each example assessment:

    a. Calculate the **absolute difference** between the marker's score and the reference score

    b. Calculate **how wrong it is possible to be**, that is, the greater of the difference between the reference score and 0 or the reference score and 100

    c. Calculate the **raw accuracy** as 1 - (a / b)

2. Calculate the **mean** of the set of raw accuracy scores.

3. Calculate the **absolute deviation (AD)** of the set of raw accuracy scores. If less than 0.01, consider it equal to zero.

4. Calculate the **adjusted mean (AM)** as mean • (1 - AD). This is the mean, reduced by a level reflecting how accurate the marker is.

5. Calculate the **deviation factor (DF)** according to the table above, using the *consistency* setting.

6. Adjust the deviation factor according to the marker's score. This is represented by the diagram above in fig. 2. The slopes are calculated as follows:

    a. Calculate the slope according to the Curves table, using the *consistency* setting.

    b. Calculate 2^a

c. Multiple the deviation factor by (b • mean - b + 1). This equation gives the graph in fig. 2.

d. If the deviation factor is less than zero, set the deviation factor to zero.

7. The **calibrated score (CS)** is given by AM • DF + mean • (1 - DF), restricted to the range 0–1.

8. Finally, the score is scaled according to the curve specified in the *comparison* setting. Up until now we had assumed that the curve was linear, i.e. that the comparison setting was the default, 5.

a. Calculate the **grading curve** according to the table above, using the *comparison* setting.

b. If $a \geq 1$, the calibrated score is given by $1 - (1 - CS)^a$

c. If $a < 1$, the calibrated score is given by $(CS)^{1/a}$

d. These equations give the curves in fig. 1 above.