Motivating Formative Assessment Engagement for Flipped Instruction Content

SoTL Commons 2019
Kurt Schmitz

Abstract

Hybrid and Blended instructional approaches are increasingly being applied to University level curriculum. These techniques shift a significant burden for engaging course content to the students as an outside-the-classroom activity. This study of contrasting hybrid classrooms (two medium size and two large size) investigates the efficacy of completion scores in an open gradebook to motivate student engagement with formative assessments. Definitions are provided for hybrid instruction, and formative assessment to establish clear measures for data collection in this field experiment. A between-groups ANOVA analysis is presented with conclusions and recommendations.
Agenda

• Instructional Concepts related to Flipped Instruction
  ◆ Flipped Instruction
  ◆ Summative/Formative assessments
  ◆ Completion/participation scores
  ◆ Transparent gradebooks
• Research Question:
  *Should flipped instruction formative assessments be recorded in your grade-book as completion scores?*
• Study Methodology
  ◆ Field Experiment Concepts
  ◆ Design
• Results

Definitions: Inverted / Flipped Instruction

• Inverted Instruction:
  “Events that have traditionally taken place inside the classroom now take place outside the classroom and vise versa.” (Lage 2000)

• Flipped Classroom Instruction:
  “The flipped classroom model involves courses that move the traditional lecture, or content dissemination, away from face-to-face hours and into online delivery outside of class time. The face-to-face class time is used for practice and actual application rather than for introducing the content being studied.” (Hill 2012)
  “Most research on the flipped classroom employs group-based interactive learning activities inside the classroom” (Bishop & Verleger 2013)
  “An educational technique that consists of two parts: interactive group learning activities inside the classroom, and directed computer-based individual instruction outside the classroom.” (Bishop & Verleger 2013)
Assumptions & Success Factors

“Flipping the classroom assumes students will take control of their learning in terms of the pace of study, mastery of content, and responsibility for coming to class prepared.” (Davies et al. 2013)

- Factors Influencing Learning Outcomes for Flipped Instruction (Lim & Morris 2009)
  - Age
  - Prior Experience
  - Delivery format preferences
  - Average study time
Challenge: Self-Directed study

- Online learning has a history of high drop-out rates. (Lee & Choi 2011)
- Not just MOOCs, but also Online courses at Universities (Levy 2007) and Community Colleges (Simon et al. 2007)
- Self-Directed Learning starts with motivation.
  “Motivation plays a very significant role in the initiation and maintenance of effort toward learning and the achievement of cognitive goals.” (Garrison 1997)
  “lack of motivation is the major reason for student drop-outs in online course.” (Kim 2004)

Tools for motivating Self-Directed Learning

Formative Assessments
- Formative assessment feedback facilitates motivation and “deep learning”. (Higgins et al. 2002)
- Formative assessment is used by students to adjust their learning tactics (Cauley & McMillan 2010)

Online Gradebook
- “Obtaining good grades is an overarching concern... Students made it quite clear that all other goals were secondary.” (Pressley et al. 1998)
- “knowing their progress and grades in the class gave them a sense of satisfaction and motivation.” (Docan 2006)
Concept: **Summative & Formative Assessments**

“When the cook tastes the soup, that’s formative; when the guests taste the soup, that’s summative.”
Bob Stake (in Scriven 1981)

Formative tests allow the student to ask
“How am I doing?”

With summative assessments it is often
“How did I do?”

---

**Definitions: FA & SA**

**Assessment**
- Making a judgement according to standards, goals & criteria.

**FA: Formative Assessment/Evaluation**
- Conducted *during* the development/improvement
- Requires feedback revealing a ‘gap’ between actual & standard
- Requires indication of how to improve, to reach the standard

**SA: Summative Assessment/Evaluation**
- Conducted *after* completion for the benefit of an external decision maker (e.g., the teacher)
- Involve *evidence* of student achievement
- Typically involves an external evaluator for credibility.
### Examples: FA & SA

<table>
<thead>
<tr>
<th>Summative (SA)</th>
<th>Formative (FA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Pop quizzes</td>
<td>• Practice Questions</td>
</tr>
<tr>
<td>• End of activity scored quizzes</td>
<td>(golfing without a scorecard)</td>
</tr>
<tr>
<td>• Tests &amp; Exams</td>
<td>• Formative use of summative tests</td>
</tr>
<tr>
<td>• Graded assignments</td>
<td>(golfing and hiding your scorecard)</td>
</tr>
<tr>
<td>◦ Essays</td>
<td>• Prototypes</td>
</tr>
<tr>
<td>◦ Problem solving</td>
<td>• Comment-only marking</td>
</tr>
<tr>
<td></td>
<td>• Peer &amp; self-assessment</td>
</tr>
</tbody>
</table>

For the purpose of this study it is stipulated that FA are an effective component of STEM instruction. The role of FA in learning and cognitive development has been widely studied. (Black & Wiliam 1998; Gikandi et al. 2011)

### Concept: Completion/Participation Grades

**Participation Grades:**
A score in the gradebook for “doing” a task regardless of quality & performance.

**Completion Grade:**
A score in the gradebook for “doing all” of a task, regardless of quality & performance.
Participation: It’s not all good

• “Participation” does not lead to higher grade performance
  ◆ Online Interaction using discussion boards (Davies & Graff 2005)
  ◆ Peer & Self-assessed “participation” in formal grading procedures does not help (Gopinath 1999)

• Extrinsic Rewards (Grades) can undermine intrinsic motivation (Deci, et al. 2001) and interfere with the process and quality of learning and also reduce enjoyment for the topic. (Docan 2006).

Research Question

*Should flipped instruction formative assessments be recorded in your grade-book as completion scores?*

• Does the explicit link to grading make a difference in student’s completing formative assessments?

• Does the explicit link to grading make a difference in student test/exam performance?
SPOILER: RESULTS

- Formative Assessment Engagement is a statistically significant predictor of Test performance ($p$-value $\leq 0.000$, $R^2$ between 0.75 and 0.94) [Replication finding & validity check]

- Completion rewards increase the FA Engagement quantity ($p$-value $\leq 0.000$).

- Completion rewards do NOT increase Unit Test performance ($p$-value $>0.10$). Results suggest quality suffers.

Treatment

Formative Assessment

- “Practice questions” where accuracy is not recorded/tracked.
- Allows students to practice knowledge recall.
- Students receive immediate “feedback” (question level gap & guidance as well as overall progress)

Completion Scores

- Recording a score into the Open Gradebook when the Formative Assessment is “complete”.

- If students are “chasing scores”, then any missing/zero mark should motivate students to engage & complete the formative assessments.

- If formative assessments are effective, then increasing student engagement with FA will increase Test/Exam performance.
Example: Formative Assessment

Example: Online Gradebook Completion Scores

- Zero score – not completed by due date
- Score - % complete by due date
- Missing score – something due in the future
**Hypothesis**

**H1:** Formative Assessment Engagement is positively related to increased performance on Test/Exams.

**H2:** Recording and exposing a completion score for formative assessments in an open gradebook will increase completion of formative assessments.

**H3:** Recording and exposing a completion score for formative assessments in an open gradebook will increase performance on Test/Exams.

**Study Methodology**

- Longitudinal Field Experiment
- Treatment & Control
- Independent Measures vs. Repeated Measures
- Other Design considerations
  - Carry-over effects & wash-out
Longitudinal Field Experiment

Longitudinal Study Group
- T1: Measure baseline
  Unit Test #1 prior to treatment
- T2: Treatment
  Expose completion scores in online gradebook (3 weeks)
- T3: Measure response
  Unit Test #2 after treatment

Control Group
- T1: Measure baseline
  Unit Test #1 prior to treatment
- T2: no treatment
  Lack of treatment (3 weeks)
- T3: Measure response
  Unit Test #2 w/o treatment

Independent Measures vs. Repeated Measures

Independent Measures (Between Groups Study)
- Multiple groups (Treatment & Control)

Repeated Measures (within groups study)
- Multiple groups
- Cross-over treatment
**Study Design: 4 cohorts †**

Medium size classes (45 enrolled)
- Treatment group (n=27)
- Control group (n=37)

Large size classes (300 enrolled)
- Treatment group (n=228)
- Control group (n=273)

† Only students who took the final exam, and agreed to participate in study are included.

---

**H1: FA Engagement increases Exam performance**

- **OLS Regression:** \( UTn = \beta * FAEn \)
- Actual Formative Assessment Engagement (_FAEn_) measured as % completion of practice questions prior to Exam.
- FAE is concurrent with Treatment (recording completion scores in the Online gradebook ... or not recording scores in the online gradebook.)
H1: FA Engagement → Exam performance?

<table>
<thead>
<tr>
<th>Panel</th>
<th>N</th>
<th>UT1: Baseline</th>
<th>UT2: Retest 1</th>
<th>UT3: Retest 2</th>
<th>UT4: Retest 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>64</td>
<td>μ Exam score 79.81</td>
<td>β=9.575***  R²=0.79</td>
<td>μ Exam 85.13</td>
<td>β=9.292***  R²=0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
<td>μ Exam 80.36</td>
<td>β=8.898***  R²=0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>μ Exam 86.04</td>
<td>β=9.348***  R²=0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
</tr>
<tr>
<td>Exp 1</td>
<td>27</td>
<td>μ Exam 77.48</td>
<td>β=9.328***  R²=0.75</td>
<td>μ Exam 82.37</td>
<td>β=9.121***  R²=0.91</td>
</tr>
<tr>
<td>Group 1</td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
<td>μ Exam 80.67</td>
<td>β=8.808***  R²=0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>μ Exam 87.16</td>
<td>β=10.11***  R²=0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
</tr>
<tr>
<td>Exp 1</td>
<td>37</td>
<td>μ Exam 81.51</td>
<td>β=9.064***  R²=0.85</td>
<td>μ Exam 87.14</td>
<td>β=9.411***  R²=0.91</td>
</tr>
<tr>
<td>Group 2</td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
<td>μ Exam 80.14</td>
<td>β=8.967***  R²=0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>μ Exam 85.23</td>
<td>β=8.911***  R²=0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
</tr>
<tr>
<td>Experiment 2</td>
<td>501</td>
<td>μ Exam 85.13</td>
<td>β=9.035***  R²=0.86</td>
<td>μ Exam score 81.80</td>
<td>β=9.703***  R²=0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
<td>μ Exam score 67.02</td>
<td>β=7.314***  R²=0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>μ Exam score 72.02</td>
<td>β=7.660***  R²=0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
</tr>
<tr>
<td>Exp 2</td>
<td>228</td>
<td>μ Exam score 77.00</td>
<td>β=8.552***  R²=0.87</td>
<td>μ Exam 81.28</td>
<td>β=8.573***  R²=0.92</td>
</tr>
<tr>
<td>Group 3</td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
<td>μ Exam 68.93</td>
<td>β=7.441***  R²=0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>μ Exam 72.85</td>
<td>β=8.546***  R²=0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
</tr>
<tr>
<td>Exp 2</td>
<td>273</td>
<td>μ Exam score 75.83</td>
<td>β=8.414***  R²=0.86</td>
<td>μ Exam 82.25</td>
<td>β=8.819***  R²=0.90</td>
</tr>
<tr>
<td>Group 4</td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
<td>μ Exam 65.43</td>
<td>β=7.202***  R²=0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>μ Exam 71.33</td>
<td>β=7.829***  R²=0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pvalue &lt;0.000</td>
<td></td>
</tr>
</tbody>
</table>

H1 Conclusion

- FA Engagement is statistically significant predictor of Unit Test scores.
- $R^2$ between 0.75 and 0.94.
- Relationship holds for all Sessions (Term/Semester) and for all Groups (Sections with different treatments) and for all test events.

Support H1

Formative Assessments engagement is associated with increasing Test Scores.
H2: Participation/Completion Scores increase *quantity* of FA Engagement

- OLS: $\text{FAE}_n = \beta PANNEL$
- Treatment: Recording completion grades into Online Gradebook.
- Groups 1 & 3 receive treatment after UT1 and before UT2.
- Groups 2 & 4 receive treatment after UT2 and before UT3.
- Treatment = recording completion scores in the Online gradebook (default is not recording scores in the online gradebook.)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>T1</th>
<th>FAe1 Mean (StDev)</th>
<th>T2</th>
<th>FAe2 Mean (StDev)</th>
<th>$\Delta$ FAe T2-T1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp 1 Group 1</td>
<td>27</td>
<td>n/a</td>
<td>62.32 (40.05)</td>
<td>✓</td>
<td>82.83 (27.14)</td>
<td>+20.51</td>
</tr>
<tr>
<td>Exp 1 Group 2</td>
<td>37</td>
<td>n/a</td>
<td>75.94 (35.52)</td>
<td>n/a</td>
<td>84.78 (26.91)</td>
<td>+8.84</td>
</tr>
<tr>
<td><strong>ANOVA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\Delta$ FAe = $\beta T2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$F=12.05$, $P$-value&lt;0.000</td>
</tr>
<tr>
<td><strong>Experiment 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp 2 Group 3</td>
<td>228</td>
<td>n/a</td>
<td>78.70 (32.17)</td>
<td>✓</td>
<td>88.96 (25.17)</td>
<td>+10.26</td>
</tr>
<tr>
<td>Exp2 Group 4</td>
<td>273</td>
<td>n/a</td>
<td>78.57 (32.04)</td>
<td>n/a</td>
<td>84.33 (29.31)</td>
<td>+5.78</td>
</tr>
<tr>
<td><strong>ANOVA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\Delta$ FAe = $\beta T2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$F=25.38$, $P$-value&lt;0.000</td>
</tr>
</tbody>
</table>
H2: Completion Scores → FA Engagement (quantity)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>T3</th>
<th>FAe3 Mean (StDev)</th>
<th>Δ FAe T3 - T1</th>
<th>T4</th>
<th>FAe4 Mean (StDev)</th>
<th>Δ FAe T4 - T1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1</td>
<td>64</td>
<td>85.37 (21.49)</td>
<td>+15.17</td>
<td>83.96 (26.81)</td>
<td>+13.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1 Grp 1</td>
<td>27</td>
<td>✓</td>
<td>86.98 (20.26)</td>
<td>+24.65</td>
<td>✓</td>
<td>82.83 (27.14)</td>
<td>+15.08</td>
</tr>
<tr>
<td>S1 Grp 2</td>
<td>37</td>
<td>✓</td>
<td>84.20 (21.49)</td>
<td>+8.26</td>
<td>✓</td>
<td>84.78 (26.91)</td>
<td>+12.35</td>
</tr>
<tr>
<td>ANOVA</td>
<td></td>
<td>ΔFAe = β*T2</td>
<td>F=13.90 P-value&lt;0.000</td>
<td>ΔFAe = β*T2</td>
<td>F=4.22 P-value=0.044</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 2</td>
<td>501</td>
<td>86.12 (25.27)</td>
<td>+7.49</td>
<td>87.64 (27.58)</td>
<td>+8.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2 Grp 3</td>
<td>228</td>
<td>✓</td>
<td>88.07 (22.23)</td>
<td>+9.37</td>
<td>✓</td>
<td>88.91 (25.17)</td>
<td>+10.21</td>
</tr>
<tr>
<td>S2 Grp 4</td>
<td>273</td>
<td>✓</td>
<td>84.49 (27.49)</td>
<td>+5.92</td>
<td>✓</td>
<td>86.24 (29.31)</td>
<td>+7.64</td>
</tr>
<tr>
<td>ANOVA</td>
<td></td>
<td>ΔFAe = β*T2</td>
<td>F=19.41 P-value&lt;0.000</td>
<td>ΔFAe = β*T2</td>
<td>F=20.39 P-value&lt;0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

H2 Conclusions

- FA engagement (FAe) increases for groups that receive the treatment (completion score rewards).
- The advantage is persistent, but degrades over time.

Accept H2

Completion score rewards increase FA engagement (quantity)
**H3: Do Completion Scores for Formative Assessments improve Exam Scores?**

- **ANOVA:** \( \text{Performance}_{-}\text{Change}(\text{UT}_n-\text{UT}_1) = \beta \times \text{PANNEL} \)
- **Treatment:** Recording completion grades into Online Gradebook.
- Unit Test 1 is a Baseline (no groups receive the treatment)
- UT2 is after one group receives the treatment
- UT3 is after both groups receive the treatment
- UT4 measures delayed effects.

### Results H3

<table>
<thead>
<tr>
<th></th>
<th>UT1</th>
<th>UT2</th>
<th>UT2-UT1</th>
<th>UT3</th>
<th>UT3-UT1</th>
<th>UT4</th>
<th>UT4-UT1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp 1</td>
<td>79.81</td>
<td>85.13</td>
<td>+5.32</td>
<td>80.36</td>
<td>+0.54</td>
<td>86.04</td>
<td>6.23</td>
</tr>
<tr>
<td>Grp 1</td>
<td>77.48</td>
<td>82.37</td>
<td>+4.89</td>
<td>80.67</td>
<td>+3.19</td>
<td>87.16</td>
<td>9.68</td>
</tr>
<tr>
<td>Grp 2</td>
<td>81.51</td>
<td>87.14</td>
<td>5.62</td>
<td>80.14</td>
<td>-1.38</td>
<td>85.23</td>
<td>3.71</td>
</tr>
<tr>
<td>ANOVA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\text{UT}_n-\text{UT}_1 = \beta \times \text{T2})</td>
<td>(F=0.05, p=0.82)</td>
<td>(F=1.28, p=0.26)</td>
<td>(F=2.75, P=0.10)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp 2</td>
<td>76.36</td>
<td>81.80</td>
<td>+5.44</td>
<td>67.02</td>
<td>-9.34</td>
<td>72.02</td>
<td>-4.34</td>
</tr>
<tr>
<td>Grp 3</td>
<td>77.00</td>
<td>81.28</td>
<td>+4.28</td>
<td>68.93</td>
<td>-8.07</td>
<td>72.85</td>
<td>-4.15</td>
</tr>
<tr>
<td>Grp 4</td>
<td>75.83</td>
<td>82.25</td>
<td>+6.42</td>
<td>95.43</td>
<td>-10.4</td>
<td>71.33</td>
<td>-4.50</td>
</tr>
<tr>
<td>ANOVA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\text{UT}_n-\text{UT}_1 = \beta \times \text{T2})</td>
<td>(F=2.79, p=0.09)</td>
<td>(F=2.31, p=0.13)</td>
<td>(F=0.11, P=0.74)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
H3: Medium Size Class (Analysis)

- Baseline T1 Gap is performance gap between groups (between Treatment group and Control group).

- UT2 Gap is performance gap after treatment and Retest: Gap got bigger. The treatment has a marginal negative impact on performance! (not statistically significant)

- UT3 Gap is performance gap after both groups receive treatment: Gap got smaller. The treatment benefits may delayed by a testing cycle (~3 weeks).

H3: Medium Size Class (Analysis)

- Performance Gap after UT4 is migrating in direction of baseline. If there is a delayed benefit (~3 weeks or 1 test cycle), then both groups are now getting the effects of the treatment.

- ANOVA suggests that any performance effects are marginal (not statistically significant for the Medium size classes).
H3: Large Size Class (Analysis)

• Baseline T1 Gap is performance gap between groups (Treatment group G3 has a performance advantage on the Control group G4)

• UT2 Gap is performance gap after treatment and Retest: Gap flipped. The treatment appears to have a negative impact on performance (but not statistically significant).

• UT3 Gap is performance gap after both groups receive treatment: Gap flipped again. The treatment effect may be delayed by a testing cycle (~3 weeks).

H3: Large Size Class (Analysis)

• Performance Gap after UT4 is migrating toward the baseline. If there is a delayed effect (~3 weeks), then both groups are now getting the effects of the treatment.

• ANOVA suggests that any performance effects are marginal (not statistically significant for the Large size classes).
Conclusions

• Formative Assessment engagement has a statistically significant positive effect on Test/Exam performance.

• Completion/Participation scores in an Open Gradebook increase Formative Assessment engagement (quantity).

• Completion/Participation scores in an Open Gradebook have an insignificant effect of Unit Test performance.
  ◆ Effects may initially be negative.
  ◆ Positive effects may be delayed by a unit test cycle (~3 weeks).