Assessing with essays: Sources of error and mitigating responses

Symposium on Tertiary Assessment and Higher Education Student Outcomes
17-19 November, 2008; Wellington, NZ
Gavin T L Brown, PhD
University of Auckland
Goal of Grading

- Grade represents well the value of student performances (Accuracy; Truthfulness)
- Grade would be accepted by other members of a community of experts working in a discipline as an accurate representation of the value of student work (Consistency, Consensus)
- Community of grade users accept that grades represent accurately the value of student performances
- RELIABILITY of grades is a requirement for their utility
Confusion in what grades mean


• What does grade C mean?
  – Parents⇒ average performance, acceptable, no problems, my child is doing ok
  – Teachers⇒ lowest grade, unacceptable performance, warning that behaviour is problematic, insufficient learning taking place

• So what do grades at a university mean?
• Consider 3 options...
What do grades mean?

1. Place in rank order
   - Top few
   - Middle some
   - Bottom many

2. Amount answered correctly
   - 85%+ correct
   - 65-84% correct
   - 50-64% correct

3. Quality of performance
   - Excellent;
   - Good;
   - Acceptable;

Are these interchangeable notions?
Rank Order: Norm-Referenced Interpretation

- **Well established:** Instructors can order people by competence, ability, performance
- **Assumption:** Position is relatively stable;
- **Consequence:** Rank resists instruction, so why bother?
Problem with Norm-Reference

• We normalise on our own population.
  – The best in my class at my university must be as good as the whole population of university students in the country (and the world)
  – The worst in my cohort must be really bad
Problem with Norm-Reference

- BUT Rank is independent of actual quality
  - It depends on who else is in the group not actual ability.
  - Bad in a strong group might be good
  - Good in a weak group might be poor
- School teachers not good at knowing what national norms are
  - Perhaps university lecturers are better?
Percentage Correct

• Long-Standing Conventions
  – The task is scoreable at the discrete level
  – All items worth the same mark are equally difficult and fully interchangeable
  – Adding up the number of items correct tells you how much & how well someone has learned the material
  – Getting more than ½ right is minimum requirement for passing
Problems with Percentage Correct

- Not all tasks can be scored discretely
  - They don’t consist of multiple parts/bits that can be counted
  - Such as...
    - Essays
    - Course work assignments
    - Projects
    - Discussions, Presentations/Forums
    - Scrap book
    - Portfolio
    - Production of ICT application
    - Critical Comments/Reviews
    - Reflective writing
    - Report writing
Problems with Percentage Correct

• Items in tests are NOT equally difficult
  – Should not have equal weighting towards total
  – Classical Test theory is a lie;
  – Item response theory needed to adjust score by difficulty
  – Getting different items right indicates different strengths but sum score is same → misleading interpretations
    • Hence the NCEA system attempts to identify components of performance through standards
Problems with Percentage Correct

• 51% is NOT magical pass
  – Unacceptable if easy test;
  – Highly admirable if test is difficult
  – High correct needed for professions
    • % correct for landing a plane should be 100 otherwise you shouldn’t get on!!
    • % correct for surgery should be 99.99 otherwise you shouldn’t let the cutting begin

• Hence the % correct needed for each grade should be set by a process that takes into account the difficulty of the items and the expectations of the community for those certified to work in a domain
Mixing % correct with Rank Order

• aka grading on the bell curve
• Assigning a grade based on performance relative to others
  – Tests may be (by mistake) very hard or very easy
  – So place students in rank order
  – Move the distribution up or down to centre the normal bell-shaped distribution on 50%
Mixing % correct with Rank Order

- Award percentage grades based on position in distribution
- Really poor result ➔ everyone moves up
- Really good result ➔ everyone moves down

• BUT the final grade/score is a LIE!!!
  • We only know relative position

• No inherent reason to deny many a high grade; or to reward the best people with a high grade if their work is not high quality
Grades as standards

- The communal judgement of a group of peers
- People who use our grades (employers, parents, students) MUST have confidence that we are competent to judge and that we judge competently
  - Competence implies validity and reliability
Grading according to quality

- Natural & Normal process of classification
  - Degrees of quality;
  - frequently 3 levels; good, okay, bad;
    - everything else is a variant
- These classificatory judgements are commonplace in academia
- For example....
Standards in Education: Reviewing Research Publications

- Published Article
  - Accept as is (rare)
  - Accept with minor changes (sometimes)
    - Resubmission invited after major changes, likely to be accepted (usual)
      - Reject will not be accepted (common)
Standards in Education: Thesis Grades

Pass: Accepted as is (few)

Pass: Minor revisions approved by supervisor (most)

Fail: Major revisions required, re-enrol and submit in 6-12 months (few)

Fail: Unlikely to pass, withdraw (rare)
Standards in Education: Academic Grades

- Excellent; Highly Competent; \((A)\)
- Good; Competent \((B)\)
- Adequate; Minimally Competent \((C)\)
- Inadequate; Not Competent \((D)\)

Grades for essays represent holistic quality judgements; NOT rank order in class; NOT number of points made correctly
Difficulty & Standards

- Expectations vary by previous learning
  - publishing researcher > doctoral candidate
  > masters student > final year undergraduate
  > first year undergraduate

- Expectations vary by nature of course
  - Advanced course > intermediate course
  > introductory course

- SO grades must be set relative to commonly understood notions of quality of performance expected for each grade at each level of progression through university
Essays

- long time—some 3,000 years ago in Imperial China
- The core is the task, prompt, or question
  - instructs the student as to the type of writing they are expected to engage in.
  - cognitive task
    - discuss, compare, contrast, or analyse, etc.
  - content
    - the causes of World War I, the impact of the setting on a character’s development, or the role of mutation in disease)
  - Linguistic task IMPLICIT
    - Range & richness of language; organisational structure; accuracy of grammar, spelling, punctuation; amount written
Why used

• usually crafted quite quickly
• clearly focuses attention on important cognitive and curriculum objectives.
  – Integrated thought
  – Handling of complex data
  – Sophistication of reasoning
  – Comprehensiveness of data coverage, etc.
• The real work in an essay is in the scoring
Essay Examinations

- a set period of time,
- a cogent response to a task or prompt
- not previously seen, and
- produce on-demand at a certain time and place (there is no going away to look things up and finish tomorrow).
- a first draft piece of writing

HENCE: students generally do not produce their best work
- So a good reason to include take-home essays as part of overall grading
Issues in scoring essays

• Do we score consistently enough to warrant high-stakes use of essay scores?
  – What techniques for determining consistency/consensus?

• How can sources of any inconsistency be identified?
  – what are they? How serious are they?

• What can be done to mitigate those sources?
Estimating Levels of commonality

- Consensus among judges
  - Degree to which judges give the same grade
  - Exact & approximate
  - Exactly same minimum recommended = 70%;
  - approximate (+/-1 sub-grade) 90%

- Consistency among judges
  - Giving same relative order among essays even if grades not similar
  - My high might ≠ your high; but we distribute higher and lower grades in a similar way
  - At least .70 for overall grading purposes
Can essays be scored consistently?

- Consider ETS Advanced Placement examinations
  - End of HS entry to university in the USA
  - high-stakes, voluntary examinations at end of specialised courses
  - 2 components: machine scores + human scored
  - Reliability coefficients computed among judges

- moderate reliability for
  - language, literature, humanities subjects
# AP Reliability 2003

<table>
<thead>
<tr>
<th>Subject</th>
<th>Free-Response</th>
<th>Multiple-Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Lang &amp; Composition</td>
<td>.67</td>
<td>.88</td>
</tr>
<tr>
<td>English Lit &amp; Composition</td>
<td>.70</td>
<td>.86</td>
</tr>
<tr>
<td>French Lang</td>
<td>.85</td>
<td>.92</td>
</tr>
<tr>
<td>French Lit</td>
<td>.75</td>
<td>.89</td>
</tr>
<tr>
<td>International English Lang</td>
<td>.84</td>
<td>.88</td>
</tr>
<tr>
<td>Spanish Lang</td>
<td>.67</td>
<td>.86</td>
</tr>
<tr>
<td>Spanish Lit</td>
<td>.75</td>
<td>.85</td>
</tr>
<tr>
<td>US History</td>
<td>.65</td>
<td>.91</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>.74</strong></td>
<td><strong>.88</strong></td>
</tr>
</tbody>
</table>

Consistency in essay scoring

- NOT unique to AP: Many other high-stakes testing systems that use essays to test written language and content knowledge find similar levels of consistency
- Correlation of $r=0.70$ between markers seems to be the absolute minimum threshold for trusting grades in high-stakes settings
Consistency in essay scoring

- ELLIS PAGE study of 1,000 scripts, 6 judges and a computer scoring system
  - Computer Rating used:
    - Intrinsic variables: grammar accuracy, vocabulary
    - Approximations: length, ratio of active to passive voice
  - Average correlation of agreement between Computer and Humans HIGHER than Humans with Humans

- HENCE, essays have a high proportion of language skill and ability in their scoring
  - This needs to be considered especially in design of essay tasks
Sources of variability/unreliability

• Using generalisability theory
  – Analysis of variance according to components involved in an essay scoring event
    • Main effects
      – Student; task; marker;
    • Interaction effects
      – student x task; student x marker; task x marker
  – Desirable that student effect is largest
  – Phi index of dependability >.80 suggests scores are trustworthy (variance is coming from students)
Sources of error in human judgement

- **Student Medium Impact**
  - Difference in student ability should cause difference in grades but there is random variability within students
    - extenuating circumstances at the time of the assessment can affect performance
    - We peak at different times and for different formats
    - We are inherently variable around our dominant level of ability

- **Topics/Tasks Large Impact**
  - Not all topics or tasks are equal in difficulty
    - Some topics are easier to respond to
    - Some topics have richer materials supplied
    - Some tasks have clearer instructions
    - Some tasks are better organised
  - Scores are affected by student ability to select tasks
Sources of error in human judgement

• **Marker**
  - Raters are inherently variable
    • don’t consistently pay attention to all important parts of the response
    • Markers tire, lose concentration,
    • get seduced by language instead of content

• **Marker by Topic**
  - Raters not equally able to score any task or topic
    • halo effect: liking or disliking certain topics
    • Lack of content knowledge
      - didn’t attend lecture?
      - Not my area of expertise
Sources of error in human judgement

• Marker by Student Small Impact
  – We are not usually very influenced by knowing who wrote the essay or did the task but to be sure....
    • Judge the work not the name of the student who handed it in
Mitigating Responses

- Task reorganisation
- Teach organisational skills
- Scoring Rubrics
- Exemplars
- Marker Behaviour
- Monitor reliability
Mitigating responses to error: Reorganise Tasks

- Reduce emphasis on rhetorical style or structure.
  - Example: from a stage 1 essay topic.

Evaluate the sample items from the XYZ test of intelligence.

What are the intelligence factors assessed by these items?
  - Identify the mental ability that each item type is testing.
  - Explain why the item types group into the factors you have chosen.
  - What labels from Carroll’s taxonomy best describe your factors?
  - What theorist or theory is most associated with the factor pattern you have chosen?
  - Why are these factors important measures of intelligence?

Explain what kind of relationship you would expect there to be between the factors you have identified.
  - What kind of correlation and/or factor pattern will there be?
  - What kind of hierarchy, if any, will there be?
  - What does this relationship pattern say about the nature of intelligence as measured by the XYZ test?
Mitigating responses to error: Teach Organisational Skills

- Marshall (SLC at University of Auckland)
  - Taught how to write introductions to academic essays
  - Students improved substantially in 2 hour training session
  - Introduction score at end of training correlated substantially with total essay grade assigned later in semester by a tutor ($r=.59$)
  - Teach them and they will learn...
Mitigating responses to error: Develop & Use Scoring Rubrics

- Clear marking guidelines that identify key criteria and indicators for each grade
- Can be developed with 2 approaches
  - **EBB**: sort essays into 4 groups (D, C, B, A); discuss what all the C have that none of the D have; set a rule for determining difference; repeat for all boundaries
  - **Bookmark**: create sets of essays that are in ascending order of quality; grade individually; compare & discuss reasons; re-rate; repeat with 2nd set; use reasons to create rubrics for each grade
Mitigating responses to error: Exemplify Quality

- Select essays that have been scored at each grade level
- Annotate each essay with explanations for why it got its grade
- Make these available to tutors, markers, and students
Mitigating responses to error: Improve marker scoring

- Use a pre-specified marking rubric
- Use at least two markers per script
- Score in a suitable non-distracting environment
- Mark all of one topic before moving on to the next
- Mark all of the same topic in one sitting, and if not possibly re-calibrate
- Recalibrate by remarking some essays blind (marks removed) and re-learn marking standards
Mitigating responses to error: Topic or Task Design

- Give everyone the same essay topics
  - No choices
- Require each student to complete at least 3 topics in an examination
- Require at least 2 or 3 topics as part of course-work
- Check the clarity of topic instructions and task requirements
Mitigating responses to error: Monitor Reliability of Scoring

- Calculate % identical scores (aim for >70%)
- Calculate correlation of scores (aim for > .70)
- Report reliability statistics in examiner reports
- Use the reliability information in determining whether adjustments need to be made at the boundary for D—C; B—A
  - That is: given the error in scoring is it really true that the D students failed? And likewise are there really more A students than you have reported?
Conclusion

- The high-stakes use of essays
  - Problematic because of difficulty in obtaining reliable scores
- We know a lot about what causes error and have developed strategies to reduce it and increase reliability
- High-stakes use requires demonstrable consistency and consensus among a community of experts who are deemed competent by peers and those hiring or teaching the students on the basis of your grades
Assessing with essays: Sources of error and mitigating responses

Symposium on Tertiary Assessment and Higher Education Student Outcomes
17-19 November, 2008; Wellington, NZ
Gavin T L Brown, PhD
University of Auckland