GRADE COMPARABILITY IN THE MATRICULATION EXAMINATION

Faculty of Behavioural Sciences/
Centre for Educational Assessment /
http://www.helsinki.fi/cea/
Marjanen Jukka/ Forum Criteriorum
WHAT IS A CRITERION?

• Merriam-Webster dictionary:
  • Criterion:
    • 1: a **standard** on which a judgment or decision may be based
    • 2: a **characterizing** mark or trait
  • Standard:
    • a **level of quality**, achievement, etc., that is considered acceptable or desirable
    • something that is very good and that is **used to make judgments about the quality of other things**
    • something set up and established by authority as a **rule for the measure** of quantity, weight, extent, value, or quality
  • Standardize:
    • to change (things) so that they are **similar and consistent** and agree with rules about what is proper and acceptable
WHY HAVE CRITERIA & STANDARDS (GENERALLY)?

1 kilogram = 2.204 pounds

• Standards are about:
  • Quality
  • Predictability
  • Consistency
  • Comparability
  • Convenience

• But also about:
  • Equity
  • Fairness
  • Transparency
  • Impartiality
  • Neutrality
WHY HAVE CRITERIA & STANDARDS IN EDUCATIONAL ASSESSMENT?

• (Formative assessment)
• Summative assessment:
  • **National tests** assess student learning on the system level:
    • Maintaining quality → If there are no evaluation criteria how can this be done? What is quality?

• **Lower secondary school leaving certificate** is used to select students to upper secondary schools:
  → To achieve a **grade 8** in mathematics a student should have to demonstrate similar mathematical competence regardless of the school they have attended or which of the school’s teachers has awarded the grade

  • From the upper secondary institutions’ point of view:
    • Predictability, consistency, convenience
  • From the students’ point of view:
    • Equity, fairness, transparency
WHY HAVE CRITERIA & STANDARDS IN EDUCATIONAL ASSESSMENT?

• **Matriculation examination** results used in tertiary institutions’ student selection
  - Since the examinations are national, the problem of grade comparability is not so much between schools or teachers…
  
  …but…

  …between the grades from different subjects.

• This will be demonstrated in this presentation both theoretically and empirically
• As a result of stratified exam selection and normative grading the grades from different subjects are not comparable:

• Notice, that an assumption of some sort of general academic competence is inherent in this line of thought!
THE USE OF MATRICULATION EXAMS IN STUDENT SELECTION: TWO EXAMPLES

Helsingin yliopisto, Matematiikka:

Metropolia, Sosiaali- ja terveysala (kevät 2014):
STATISTICAL METHODS FOR GRADE COMPARABILITY (Coe et al. 2008)

• The basic idea (simplistically):
  Compare the grades that the same candidates typically get in one subject with what they achieve in others
  → If the grades are typically/routinely lower/higher than in the other subjects they need to be adjusted
  → The methods produce "correction values" that can be added to the grades in each subject to bring them on the same scale

• Many different methods:
  • **In Finland:** Standardized marks average (SMA or SYK)
  • Aggregated subject pair analysis (ASPA)
  • Kelly’s method
  • Multilevel analysis
  • Latent trait models (Basic rationale a bit different)
AN EXAMPLE: SMA (OR SYK)

• Standardized marks average:

\[ SMA = \frac{z_{Phy} + z_{Phi} + z_{MathAdv} + \ldots}{Number\ of\ exams\ participated} \]
THE SOLUTION

• (At least theoretically) the statistical methods solve:

1. The problem of exam selection and normative grading:
   • Adjusts for differences in examinees general competence

2. The problem of how much general competence is portrayed by each grade:
   • All of these methods use some sort of an aggregate to compute the correction values (except the latent trait models…)
     → This aggregate can be thought to represent the students’ general academic competence
     → The corrected grades can be interpreted to be on the same scale as regards to the amount of general competence they portray
THE UNIDIMENSIONALITY ASSUMPTION

• If grades from different subjects are to be compared they need to measure something in common, at least to a degree (Newton 2005)

• Otherwise, we are comparing apples with oranges… or rather camels with oranges

• Grades are comparable **only** in relation to this common factor! (Coe 2010)

• Theoretically it reasonable to assume, that all exams in the matriculation examination measure something that is useful in all studies regardless of their particular discipline
  - This “something” can be called general academic competence which can be thought of as a combination of students’ thinking skills, cognitive competence, motivation, resilience and so on
Jukka Marjanen

Ylioppilasarvosanojen vertailukel poisuuongelma ja SYK-menetelmän edellytykset sen ratkaisemiseen


Ylioppilastutkintolautakunta (YTL) siirtyy ylioppilaskokeiden arvostelussa standardoitu- jen yhteispisteiden keskiarvoon (SYK) perustuvaan menetelmään. Menetelmän käyttö on mielekästä vain, jos eri oppiaineiden koemenestyksen taustalta on löydetävissä jokin yhteinen tekijä. Tässä tutkimuksessa tarkastellaan, ovatko edellytykset SYK-menetelmän käytölle olemassa. SYK-menetelmän on tarkoitus parantaa eri oppiaineista saatujen arvosan-
THE AIMS OF THE ARTICLE:

- The Finnish Board of Matriculation Examination adopted a new (statistically motivated) SYK-grading system in 2014
  - However, it has not been tested whether the competence structure that underlies students’ matriculation exam success is unidimensional (enough)

  ➔ Testing the unidimensionality assumption

- The need for the statistical methods depends on the magnitude of the difference in participants’ general competence between exams
  - If the differences are small there is no need for the complicated SMA-system

  ➔ Analyzing the general competence of the examinees in different exams
THE DATA

- The data consists of 4155 upper secondary school students’ course grades and matriculation examination results
  - Year 2012: n=2094 and year 2009: n=2061

- Course grades were averaged within subjects

- The grades in “rare” languages were combined into two variables:
  - Advanced level
  - Basic level
THE METHODS

• **Unidimensionality analysis: Confirmatory factor analysis**
  - Comparison of the fit indices for a unidimensional factor model and a bi-factor model (see fig. 1)
  - There were some missing values in the data matrix → Full information maximum likelihood estimation was used (FIML)

• **Exam selection: One sample t-tests**
  - One sample t-tests were used to analyze whether the average general competence of the participants was zero in each subject
  - Factor scores from the bi-factor model were used as a measure of general academic competence
  - If some means were negative and others positive there were significant differences in examinees competence
Figure 1a. Unidimensional confirmatory factor model

Figure 1b. Confirmatory bi-factor model with a general factor and specific factors for math/science and language subjects
## THE RESULTS: MULTIDIMENSIONALITY

### Table 1. The fit of the unidimensional and bi-factor model in 2012 and 2009

<table>
<thead>
<tr>
<th>Model</th>
<th>2012</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>chi²/df</td>
<td>TLI</td>
</tr>
<tr>
<td>Independence model</td>
<td>19356,641 / 136</td>
<td>---</td>
</tr>
<tr>
<td>One dimensional model</td>
<td>2542,63 / 119</td>
<td>0,86</td>
</tr>
<tr>
<td>Bi-factor model</td>
<td>809,44 / 107</td>
<td>0,95</td>
</tr>
</tbody>
</table>

### Table 2. The amount of course grade variance explained by the general and specific factors according to subject domain in 2012 and 2009

<table>
<thead>
<tr>
<th>Domain</th>
<th>2012</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2012</td>
<td>2009</td>
</tr>
<tr>
<td></td>
<td>2012</td>
<td>2009</td>
</tr>
<tr>
<td>General</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math/Science</td>
<td>46 %</td>
<td>40 %</td>
</tr>
<tr>
<td>Languages</td>
<td>21 %</td>
<td>20 %</td>
</tr>
<tr>
<td>**</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>**</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Languages</td>
<td></td>
<td></td>
</tr>
<tr>
<td>**</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>**</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Humanities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>**</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>**</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>All subjects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>**</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>**</td>
<td></td>
<td>**</td>
</tr>
</tbody>
</table>
### Table 3. Average competence of the examinees in different subjects matriculation examinations, general and specific competence dimensions

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics, Basic</td>
<td>-0.27</td>
<td>-0.06</td>
<td>-0.09</td>
<td>-0.30</td>
<td>-0.10</td>
<td>-0.17</td>
</tr>
<tr>
<td>Health ed.</td>
<td>-0.26</td>
<td>-0.18</td>
<td>0.00</td>
<td>-0.25</td>
<td>-0.22</td>
<td>-0.07</td>
</tr>
<tr>
<td>Geography</td>
<td>-0.13</td>
<td>0.09</td>
<td>-0.20</td>
<td>-0.18</td>
<td>-0.10</td>
<td>-0.16</td>
</tr>
<tr>
<td>Civics</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.14</td>
<td>0.02</td>
<td>-0.16</td>
<td>-0.14</td>
</tr>
<tr>
<td>History</td>
<td>0.02</td>
<td>-0.28</td>
<td>-0.25</td>
<td>-0.04</td>
<td>-0.33</td>
<td>-0.27</td>
</tr>
<tr>
<td>Psychology</td>
<td>0.11</td>
<td>-0.31</td>
<td>-0.06</td>
<td>0.11</td>
<td>-0.36</td>
<td>-0.08</td>
</tr>
<tr>
<td>Religion</td>
<td>0.16</td>
<td>-0.58</td>
<td>-0.24</td>
<td>0.13</td>
<td>-0.59</td>
<td>-0.43</td>
</tr>
<tr>
<td>Swedish, Intermed.</td>
<td>0.20</td>
<td>-0.10</td>
<td>0.33</td>
<td>0.16</td>
<td>-0.02</td>
<td>0.26</td>
</tr>
<tr>
<td>Philosophy</td>
<td>0.23</td>
<td>-0.26</td>
<td>-0.34</td>
<td>0.16</td>
<td>-0.36</td>
<td>-0.17</td>
</tr>
<tr>
<td>Physics</td>
<td>0.30</td>
<td>0.72</td>
<td>-0.18</td>
<td>0.23</td>
<td>0.77</td>
<td>0.01</td>
</tr>
<tr>
<td>Spanish, Basic</td>
<td>0.38</td>
<td>-0.04</td>
<td>0.57</td>
<td>0.40</td>
<td>0.14</td>
<td>0.58</td>
</tr>
<tr>
<td>Biology</td>
<td>0.38</td>
<td>0.04</td>
<td>-0.14</td>
<td>0.29</td>
<td>0.11</td>
<td>-0.03</td>
</tr>
<tr>
<td>Mathematics, Advanced</td>
<td>0.45</td>
<td>0.47</td>
<td>-0.08</td>
<td>0.39</td>
<td>0.44</td>
<td>0.08</td>
</tr>
<tr>
<td>Chemistry</td>
<td>0.46</td>
<td>0.54</td>
<td>-0.17</td>
<td>0.46</td>
<td>0.73</td>
<td>0.09</td>
</tr>
<tr>
<td>Swedish, Adv.</td>
<td>0.50</td>
<td>-0.04</td>
<td>0.48</td>
<td>0.52</td>
<td>-0.18</td>
<td>0.39</td>
</tr>
<tr>
<td>German, Basic</td>
<td>0.52</td>
<td>0.15</td>
<td>0.47</td>
<td>0.45</td>
<td>0.02</td>
<td>0.57</td>
</tr>
<tr>
<td>French, Basic</td>
<td>0.57</td>
<td>-0.16</td>
<td>0.57</td>
<td>0.70</td>
<td>-0.04</td>
<td>0.65</td>
</tr>
<tr>
<td>German, Adv.</td>
<td>0.66</td>
<td>0.07</td>
<td>0.66</td>
<td>0.95</td>
<td>0.26</td>
<td>0.95</td>
</tr>
</tbody>
</table>
CONCLUSIONS (UP TO THIS POINT)

• There are significant differences in participants’ general academic competence between exams
  → Some (statistical) method needs be used to improve grade comparability

• Some general competence dimension explains at least part of the grade variance in all subjects’ matriculation examinations
  → A justification for the use of statistical methods for grade comparability

• **BUT:** The competence structure underlying students’ success in the examination is multidimensional
  → This biases the results produced by the statistical methods
  → The magnitude of the bias needs to be studied further

• The actual content of the "general academic competence” cannot be determined by this data
  → Whether this construct truly is related to students’ success in tertiary education should be validated by further research
THE QUESTION

• Could all this be avoided if criterion based assessment was used instead of normative assessment?

Most of the statistical methods for grade comparability have been developed in the United Kingdom for the GCSE exams which are criterion referenced.

→ The problem of grade comparability remains:

1. Comparability **over time**
   - Are the criteria constant between the fall of 2013 and spring of 2016?
   - How well do the exam items measure the criteria in different years?

2. Comparability **between subjects**
   - What kind of a performance in mathematics is considered equal to a certain level of performance in philosophy?
     → On what basis to compare? → General academic competence
     → How to quantify? → The statistical methods

   - What level of competence in advanced English is considered equivalent to a certain level of competence in basic French?
     → Cannot be the same level on the Common European Framework of Reference for Languages
     → Instead: General academic competence
     → How to quantify? → The statistical methods
In the traditional grading system it was (theoretically) possible, that even the extremely competent students failed a test.

In the SMA (or SYK) system, it is (theoretically) possible that even the very poor students do not fail the test, if they do well in the other exams.

→ Should there be some criteria for at least passing an exam?
   - Then the problem of the equivalence of criteria between time and subjects would have to be solved.

→ Or should we just trust that these problems are only theoretical?