Assessment of individual competence: a sequential mixed model

Abstract

Aims: the assessment of individual competence in medical education is about finding a balance between having sufficient resources to make valid and reliable decisions and not using more resources than necessary. Sequential assessment, in which more resources are used for borderline performing candidates than for poorly or clearly satisfactorily performing candidates, can be used to achieve that balance. Although sequential assessment is commonly associated with larger groups of candidates to be assessed, in many practical settings numbers of candidates may be small.

Objective: this article presents a single case design with a statistical model for the assessment of individual competence that can be used regardless of the number of candidates.

Method: a worked example of a solution that can be used for an individual candidate, using simulated data in the zero-cost Open Source statistical program R version 4.0.5., is provided.

Results: the aforementioned solution provides statistics that can be used to make pass/fail decisions at the level of the individual candidate as well as to make decisions regarding the length and timing of an exam (or parts thereof) for the individual candidate.

Conclusion: the solution provided can help to reduce resources needed for assessment to a considerable extent while maximizing resources for borderline candidates. This facilitates both decision making and cost reduction in assessment.

Keywords: assessment, competence, individual, mixed model, sequential assessment.
Introduction

Apart from informing subsequent learning and practice, one of the functions of the assessment of competence in medicine and other high-stakes settings is to make valid and reliable decisions regarding progression, and longer exams are commonly associated with higher validity and reliability (1). However, given tremendous pressures on our healthcare systems, not using more resources for assessment than necessary is imperative. In addition, validity and reliability are not constant across the performance range (2). Sequential assessment, also referred to as sequential testing, can be used to address these issues and works as follows.

Performance in an assessment can range from excellent to very poor and candidates demonstrating satisfactory or better overall performance (i.e., sufficiently competent) should pass while candidates demonstrating below-standard overall performance should not pass the assessment. The standard in assessment in medical education is usually based on the concept of borderline candidate, that is: a candidate passing by the skin of their teeth, demonstrating a performance almost at the expected level with some minor issues that would not result in concerns about patient safety. The further away a candidate’s performance from that standard, the fewer resources are needed to decide whether the performance meets the standard.

Although sequential assessment is well known in the context of larger cohorts of students in an undergraduate programme, this article demonstrates that single case design methodology (e.g., (2-3)) can provide designs and statistical models for sequential assessment of individual competence regardless of the number of candidates.

Method

To illustrate the single case design solution for sequential assessment, this article uses simulated data from a hypothetical sequential assessment of clinical skills with four candidates that comprises two sits each with two parts and works as follows.

Each of the in total four parts consists of four candidate-actor encounters (i.e., stations) that assess the same five criteria that are rated in the same order across stations on the same integer scale from 1 (min) to 6 (max). Thus, we obtain five 1-6 ratings about a candidate’s performance in each station or 20 ratings per part. Although stations within each part focus on different content, the different parts are comparable in content and difficulty, and therefore an average of 3.5 over the 20 ratings in any part results in a pass. Additionally, using a mixed regression model adapted from Maric and Van der Werff (3) that accounts for both serial correlation in ratings from the same candidate (i.e., these are not 20 different candidates each obtaining one independent rating) and differences between parts and sits within candidate, we obtain a 90% confidence interval (CI) that for below-3.5 average performing candidates can be used to test $H_0: \mu = 3.5$ (i.e., the minimum score needed to pass) at $\alpha = 0.05$ against $H_1: \mu < 3.5$. If the 90% CI for a candidate includes 3.5, that candidate returns for the next part (except if we are already in the last of four parts); if the 90% does not include 3.5, the candidate in question does not return for the next part in the same sit but in the next sit. In sum, the procedure is as follows:

- 1st sit part 1: completed by all candidates;
- 1st sit part 2: candidates who failed to reach 3.5 average performance in 1st sit part 1 but had a 90% CI including 3.5;
- 2nd sit part 1: candidates who failed to reach 3.5 average performance in 1st sit part 1 and had a 90% CI excluding 3.5, plus candidates who failed to reach 3.5 average performance in 1st sit part 2 but had a 90% CI including 3.5; and
- 2nd sit, part 2: candidates who failed to reach 3.5 average performance in 2nd sit part 1 but had a 90% CI including 3.5.
All analyses were done in the *nlme* package \(^{(4)}\) in *R* version 4.0.3 \(^{(5)}\).

### Results

**Table 1** presents the ratings and average performances for the four candidates in the simulated example.

<table>
<thead>
<tr>
<th>ID</th>
<th>Sit</th>
<th>Part</th>
<th>Ratings in order provided</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>1</td>
<td>3, 3, 3, 3, 4, 3, 3, 1, 2, 3, 3, 4, 3, 4, 3, 2, 2, 2, 2</td>
<td>2.85</td>
</tr>
<tr>
<td></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>1</td>
<td>4, 2, 3, 3, 4, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>2</td>
<td>3, 4, 2, 2, 3, 4, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3</td>
<td>3.10</td>
</tr>
<tr>
<td>#2</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>1</td>
<td>2, 3, 2, 3, 3, 3, 3, 3, 4, 3, 4, 3, 3, 3, 3, 3</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>1</td>
<td>5, 4, 5, 5, 4, 5, 3, 3, 3, 3, 3, 3, 4, 3, 3, 4</td>
<td>3.90</td>
</tr>
<tr>
<td>#3</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>1</td>
<td>3, 4, 3, 3, 4, 2, 3, 3, 3, 3, 3, 3, 3, 4, 3, 3</td>
<td>3.35</td>
</tr>
<tr>
<td></td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2</td>
<td>4, 4, 5, 3, 4, 4, 3, 3, 3, 4, 4, 3, 4, 4, 4, 4</td>
<td>3.75</td>
</tr>
<tr>
<td>#4</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>1</td>
<td>5, 4, 3, 3, 5, 5, 3, 5, 5, 4, 4, 4, 4, 4, 4, 3, 3, 4, 4</td>
<td>4.00</td>
</tr>
</tbody>
</table>

To understand decision-making regarding return for next part / sit, **Table 2** presents the 90% CIs with interpretation for each of the four candidates for each part in which they participated.

<table>
<thead>
<tr>
<th>ID</th>
<th>Sit</th>
<th>Part</th>
<th>90% CI</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>1</td>
<td>[2.458; 3.288]</td>
<td>return for 2&lt;sup&gt;nd&lt;/sup&gt; sit part 1</td>
</tr>
<tr>
<td></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>1</td>
<td>[2.902; 3.502]</td>
<td>return for 2&lt;sup&gt;nd&lt;/sup&gt; sit part 2</td>
</tr>
<tr>
<td></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>2</td>
<td>[2.838; 3.358]</td>
<td>fail</td>
</tr>
<tr>
<td>#2</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>1</td>
<td>[2.731; 3.263]</td>
<td>return for 2&lt;sup&gt;nd&lt;/sup&gt; sit part 1</td>
</tr>
<tr>
<td></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>1</td>
<td>[3.571; 4.262]</td>
<td>pass</td>
</tr>
<tr>
<td>#3</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>1</td>
<td>[3.153; 3.543]</td>
<td>return for 1&lt;sup&gt;st&lt;/sup&gt; sit part 2</td>
</tr>
<tr>
<td></td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2</td>
<td>[3.549; 3.951]</td>
<td>pass</td>
</tr>
<tr>
<td>#4</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>1</td>
<td>[3.706; 4.300]</td>
<td>pass</td>
</tr>
</tbody>
</table>

Candidate #1 performed so poorly in 1<sup>st</sup> sit part 1 that a return for the 2<sup>nd</sup> sit was needed, and although the 90% CI in 2<sup>nd</sup> sit part 1 included 3.5, the 90% CI in 2<sup>nd</sup> sit part 1 was completely below 3.5, hence a fail (e.g., in an undergraduate programme with two sits per year, that could mean the candidate had to return for the same assessment next year). Figure 1 presents the mixed regression model plot for Candidate #1 to provide a visual of the model used in this article.
Candidate #2 performed so poorly in 1st sit part 1 that a return for the 2nd sit was needed, but performance in 2nd sit part 1 was good enough for part 2 in the 2nd sit not being needed. Candidate #3 needed both parts in the 1st sit but no 2nd sit, and Candidate #4 needed only 1st sit part 1.

Discussion

Some readers may wonder (i.) why we should not have all below-standard performing candidates in part 1 return for part 2, (ii.) whether we should have barely-above-standard performing candidates return for another part as well, and/or (iii.) whether we should average candidate performance across parts in the same sit instead of treating them separately. The answer is the same to all three questions: these too are options in sequential assessment (which are encountered in practice), and the setup of any assessment – sequential or not – should always be decided on in light of the context in which the assessment is to take place.

As for (i.), the rationale behind the setup chosen for this article has been to raise awareness that
there are statistical models that can help us decide not only whether we have sufficient information for satisfactorily (or higher) performing candidates to pass the assessment but equally whether performance of a struggling candidate is either such that we need more information to make a decision or is so poor that instead of allocating additional resources to assessment at this point the candidate should take more time to improve and return for assessment at a next occasion.

Regarding (ii.), the higher the stakes of an assessment, the stronger the argument in favour of having barely-above-standard performing candidates return for another part to obtain more information for decision-making, and the 90% CI resulting from the model presented in this article can in that case be used to test $H_0: \mu = 3.5$ (i.e., the minimum score needed to pass) at $\alpha = 0.05$ against $H_1: \mu > 3.5$ for any candidate with an average score above 3.5; an interval including 3.5 would then mean the candidate having to return for another part while an interval not including 3.5 would result in a pass. In the example provided, there are no candidates scoring on average above 3.5 but with a 90% CI including 3.5.

Finally, on (iii.), treating parts within the same sit as separate constitutes a way to acknowledge that assessments can be learning opportunities and hence performance may improve from one part to the next; if the two parts yield very similar performance the different statistical treatments should yield more or less the same results.

Neither for excellent performance nor for notably poor performance do we need as many resources for confident decision-making as we do for candidates whose performance is borderline. While not part of the example, it is in practice possible that even after two sits with two parts each we have a candidate who consistently had a 90% CI including 3.5; performance of such a candidate has been borderline throughout and decision-making about this candidate will be more difficult than for other candidates in this example, even after a total of four parts, and in such a case further assessment may have to be considered.

The minimum number of stations needed is to be determined in the context in which an assessment is to take place. While including large numbers of stations in each part may be too resource-intensive, too small numbers of stations per part comes will likely come at the cost of substantial gaps in content covered (reduced validity) and wide CIs (lower reliability).

Finally, although sequential assessment is commonly associated with assessments of clinical skills such as history taking, physical examination and problem solving, the concept can be applied to blocks of written knowledge assessments as well. For example, if in a set of 300 single-best answer multiple-choice questions we can create four blocks of 75 questions that are comparable in content and difficulty, candidate scoring well below or visibly above the standard in one block is likely to do so in other blocks at the time as well; we may well need a second block when dealing with a borderline candidate, but for the other two performances just mentioned that second block is probably more than what is needed.

Notes

Funding

This study did not receive financial support from external sources

Conflicts of interest disclosure

The authors declare no competing interests relevant to the content of this study.

Authors’ contributions

All the authors declare to have made substantial contributions to the conception, or design, or acquisition, or analysis, or interpretation of data; and drafting the work or revising it critically for important intellectual content; and to approve the version to be published.
Availability of data and responsibility for the results

All the authors declare to have had full access to the available data and they assume full responsibility for the integrity of these results.

References


Jimmie Leppink

PhD in Statistics Education, LLM in Forensics, Criminology and Law, and MSc in Psychology and Law from Maastricht University, the Netherlands; MSc in Statistics from Catholic University of Leuven, Belgium; currently Research Director at Hospital virtual Valdecilla (HVV), in Santander, Spain.

Mailing address

Jimmie Leppink
Hospital virtual Valdecilla
Santander, Spain; 39008