Pairwise Scaling: Did it solve the ultimate exam construction riddle of timeliness with integrity?

Marcus Edwards, BPharm, MACCP, AppDipBusProg | Australian Pharmacy Council (APC)
Beth Kerrison, BEd, MEdAdmin, MPublicPolicy | APC
Glenys Wilkinson, BSocSc, MSocWk, MAppSc, GradCertOrgCoach, FCHSM | APC
Joanna McFarlane, BA | APC

Abstract

In early 2021 our psychometric consultants suggested 'pairwise scaling' as a solution to alleviate the bottleneck of adding new items to our Intern Written exam that we trialled using innovative design, confidence in risk management and robust evaluation.

Similar to a comparative judgement technique used for marking, we designed a tool for subject matter experts (SMEs) to compare new and anchor items to produce a dataset to calculate a perceived scale rating for use as scored items in live exams.

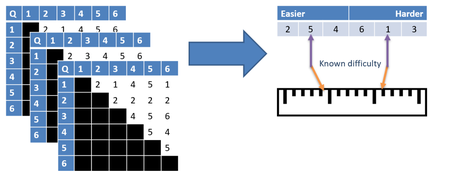
We invested time in building confidence with our SME participants in the balance between subjective and objective evaluation of an exam item which is fundamental to this approach. This goes against the technical and scientific values of your average pharmacist.

This poster displays what we have learned during our trials of this process from 2021-2023, data and analysis outputs used for evaluation, how we adapted the instructions for SMEs and what happened when we increased the quantum of questions per trial for increased data outputs.

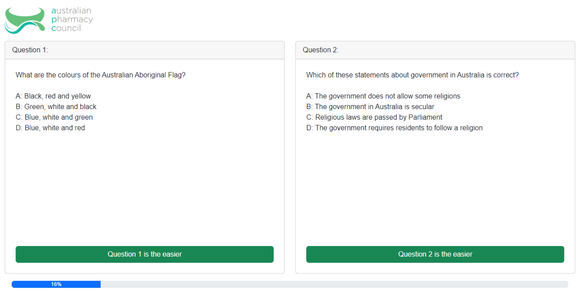
What is Pairwise Scaling?

Pairwise Scaling is a comparative judgement technique that we have tested using a unique application method to produce scale values for our exam questions before use in a live exam.

Similar to Angoff techniques for determining exam cut scores or perceived difficulty, the method relies on subject matter experts (SMEs), in our case pharmacists, that are very familiar with the exam purpose, content and competency of a minimally proficient candidate. By using anchor items in the dataset of comparison responses we can place new exam questions on our exam scale to use for scoring.



In 2021 due to ongoing effects of the COVID-19 pandemic, we were not comfortable delivering workshops face to face. We still wanted to trial this method, so we designed a custom web-based application that displays questions for comparison and collects responses:

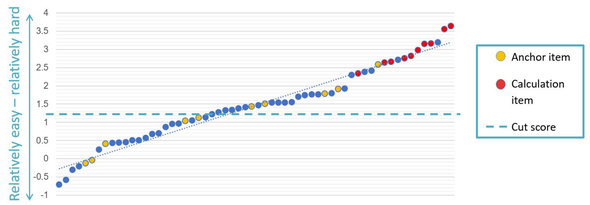


Trial 1 – October 2021

Item selection
We selected the following items to go into our first trial (new items N=50, anchor items N=10, total comparisons N=1770, SMEs N=16). Our assumption was that the SMEs would need 10 seconds per comparison. We batched questions to make the task manageable.

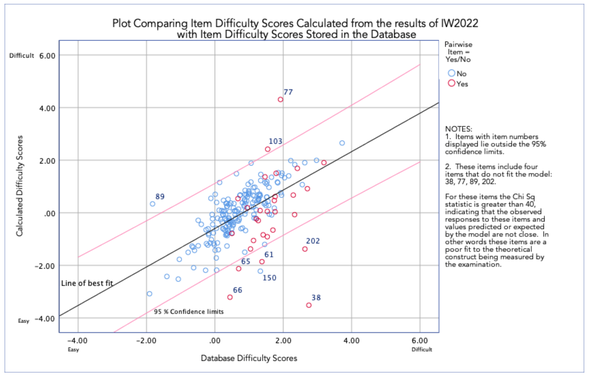
Findings
Initially, the SMEs were slower than our assumption but as their familiarity with the task increased, they became a lot faster:

Data from the workshop was analysed for judge consistency and produced perceived item difficulty for new items informed by the anchor items. Calculation questions showed up as the more difficult



Application of data
Pairwise scaling items were incorporated into our 2022 exam forms, alongside unscored questions as a safety-net for the processing, evaluation, and delivery of the Intern Written exam to generate results (Pairwise Scaling items N=32, unscored items N=32, regular items N=183).

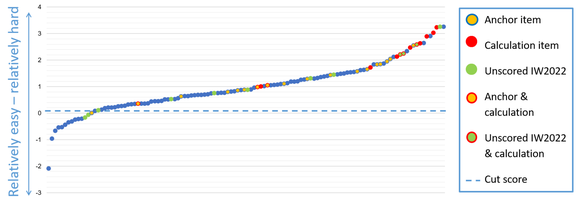
The confidence interval graph from the first session (IW2022) analysis showed consistency between pairwise scaling and live candidate data. Outliers were mostly calculation questions:



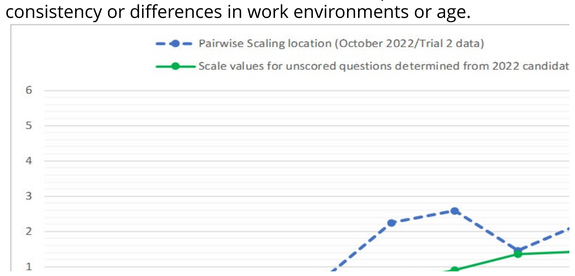
Trial 2 – October 2022

Item selection
For our second trial we used learnings from Trial 1 to adjust our training messages and tightened the range for anchor items to a narrower window around our exam cut score. We used 50% more questions overall, with a larger pool of SMEs and inbuilt a comparison with 2022 unscored items (new items N=90, unscored items from 2022 forms N=12, anchor items N-18, SMEs N=30).

Findings
Similar to Trial 1, the SMEs were faster if they had familiarity with the task and embraced our adjusted coaching. We noticed that again, calculation questions were scaled as 'difficult' in Pairwise Scaling. Data from the workshop was analysed for judge consistency and produced perceived item difficulty for new items informed by the anchor items.



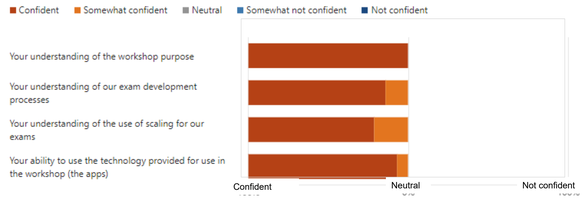
4 unscored questions were outliers and 2/30 judges were identified as outliers in trial 2 in analysis in terms of their response fits. However, there was no identified trend in response time and consistency or differences in work environments or age.



Application of data
New items used in Trial 2 were added to our 2023 exams with the perceived scale value from analysis.

SME Feedback and Conclusion

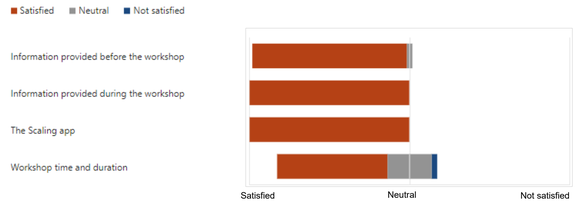
Trial 1 feedback from SMEs



We used information from the SME feedback and from trends in our analyses to tailor our training messages for Trial 2.

Conclusion (cont.)

Trial 2 feedback from SMEs



We asked the Trial 2 SMEs if they trust the process:

We are confident that by refining our training messages for SMEs and paying close attention to content area outliers that this method can add significant value to our exam construction. SMEs are capable of putting themselves in a candidate's shoes and perform this task well.

The SMEs found extended duration of Trial 2 to be overwhelming and tiresome. We have since elected to do more frequent but short pairwise sessions to ensure comfort and reliability in the process. There is some consideration allowing the SMEs to do this remotely in their own time. SME outliers identified in Trial 2 will be tracked for their suitability for the task.

The analysis between pairwise scaling and live candidate data showed that some content areas may not be conducive to the pairwise scaling methodology. Further monitoring of these outliers is required and exclusion of these content areas, with regard to the pairwise scaling, may result, if further training is not effective in reducing the outliers.

We believe our application of pairwise scaling is an effective method to alleviate the bottleneck of developing exam content for our exams. and invite all discussion and any suggestions to the process application or data evaluation and analyses at CLEAR's AEC 2023!

Acknowledgements

Consultant Psychometricians
- Professor Jim Tognolini, BEd, MEd, PhD, Director of Centre for Educational Measurement and Assessment, Director of JT Education Consulting Ltd.
- Dr Bruce Mowbray, PhD Education, BSc, MSc, JT Education Consulting Ltd.

Peter Halstead, Senior Pharmacist, Australian Pharmacy Council

Peter Robinson, MEd, BSc, DipEd, BA, Specialist Development Manager (retired), Australian Pharm.

APC Exams SME Pool