Using OpenLabyrinth for Situational Judgment Testing (SJT)

## Background

Some basic principles about SJT:

1. Situational judgment testing is best used for testing judgment, and is often used for non-clinical expertise testing. There is much more literature about it in general psychology, especially in workplace assessment.

2. There are three types of questions formats in common use in SJT in medical education at present (Patterson et al., 2012):

a: multiple response format (pick 3 out of 8)

b: ranking response format (arrange 5 responses in order, drag & drop)

c: best response format (another ‘single best answer’ format)

We mainly are using type b, the ranking response format in our SJT series in OpenLabyrinth. We see little value in type c. Type (a) will be relatively easy to score using general QUestion Rules. ‘Choose three from eight possible responses’ questions are worth 12 marks; each of the individual responses is worth 4 marks. For a simple case example, try our [‘SJT demo’ case](http://demo.openlabyrinth.ca/renderLabyrinth/index/284). (Use ‘SJTdemo’ when asked for a key.)

3. SJT is somewhat similar to Script Concordance Testing (SCT) in a few ways:

a: it tries to avoid the ‘single best answer’ concept as to which response is correct.

b: the correct response is determined by a reference panel of experts, not by what the question author considers to be correct. i.e. a baseline reference panel set of answers is needed for the set of questions.

## How SJT ranked response scoring works

For the ranking response format (drag & drop questions), players have to arrange 5 responses into an order of priority from most likely at top to least likely at the bottom. For illustration, let’s say that the responses are as follows when shown to the user: a, b, c, d, e.

And after 3 users have run the test, we get the following sets of responses:

* User 1: a, b, c, d, e;
* User 2: a, c, b, d, e;
* User 3: b, a, c, d, e;

That is, user 1 has not changed the order at all.

For this example, the exam has already been completed by a reference panel of experts. (This would be similar to SCT reports, where we compare with the reference panel scores.) When the experts’ scores are aggregated, the most popular first choice by the experts is response b, followed by a then c, d, e. That is, the experts’ opinion is that BACDE is the correct order.

So in the above, user 3 is in perfect agreement with the expert panel and would receive full marks (20/20 on that question). The 20 marks for that question are calculated as follows:

If User 3 places b as first response, and this is also the experts first response, they get 4 marks for this.

Figure 1 provides an example of how the ranking scoring system works.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ideal rank** | **Applicant rank 1** | **Applicant rank 2** | **Applicant rank 3** | **Applicant rank 4** | **Applicant rank 5** |
| B | **4** | 3 | 2 | 1 | 0 |
| A | 3 | **4** | 3 | 2 | 1 |
| C | 2 | 3 | **4** | 3 | 2 |
| D | 1 | 2 | 3 | **4** | 3 |
| E | 0 | 1 | 2 | 3 | **4** |

So going down the ideal rank column, BACDE…

User 1 scores as follows:

Response A was their first choice, rank order 1, so looking at row 2 of the table for response A, they get 3 marks (row 2, col2)

Response B was their second choice, rank order 2, so look at row 1 of the table for response B, they get 3 marks (row 1, col3)

Response C was their third choice, rank order 3, so look at row 3 of the table for response C, they get 4 marks (row 3, col4)

Response D was their fourth choice, rank order 4, so look at row 4 of the table for response D, they get 4 marks (row 4, col5)

Response E was their fifth choice, rank order 5, so look at row 5 of the table for response E, they get 4 marks (row 5, col6)

Total for user 1 = 18

User 2 scores as follows:

Response A was their first choice, rank order 1, so looking at row 2 of the table for response A, they get 3 marks (row 2, col2)

Response C was their second choice, rank order 2, so look at row 3 of the table for response C, they get 3 marks (row 3, col3)

Response B was their third choice, rank order 3, so look at row 1 of the table for response B, they get 2 marks (row 1, col4)

Response D was their fourth choice, rank order 4, so look at row 4 of the table for response D, they get 4 marks (row 4, col5)

Response E was their fifth choice, rank order 5, so look at row 5 of the table for response E, they get 4 marks (row 5, col6)

Total for user 2 = 16

User 3 scores as follows:

Response A was their first choice, rank order 1, so looking at row 2 of the table for response A, they get 3 marks (row 2, col2)

Response C was their second choice, rank order 2, so look at row 3 of the table for response C, they get 3 marks (row 3, col3)

Response B was their third choice, rank order 3, so look at row 1 of the table for response B, they get 2 marks (row 1, col4)

Response D was their fourth choice, rank order 4, so look at row 4 of the table for response D, they get 4 marks (row 4, col5)

Response E was their fifth choice, rank order 5, so look at row 5 of the table for response E, they get 4 marks (row 5, col6)

Total for user 2 = 16

The report that is generated has exactly the same layout and format as for the SCT report, so we can align spreadsheets easily with mixed format (SCT and SJT) exams.

There is no negative marking in the SJT (i.e. marks will not be deducted from your overall score for each incorrect answer given). The lowest possible mark, when the learner completely disagrees with the reference panel, is 8/20.

## How to set up an SJT Question

As noted above, we have not used the ‘pick 3 out of 8’ format of SJT questions. If you were to do so, then use the Multiple Choice format. Create your question stem as usual then add the 8 possible responses. You can then use a Counter and +4 for each correct response chosen. You would then need to create a custom report to determine the overall scores for such a scenario. We have not done this to date, as no project group has requested this.

For the Ranked Response format, we use a variant on the Drag & Drop question style.

For the question stem and context, you can either place this into the general Node text or you can place it into the stem for the SJT QUestion. It will work either way but we generally find it easier to place the stem into the Node’s text because it is then easier to see which Node and question is which, if you have to change the order of your questions.

In Script Concordance Testing (SCT) questions, it is relatively easy to create sets of Questions that can be reused. However, since the responses in SJT questions are entirely specific, you cannot generally reuse SJT questions.

There are a few important design points to keep in mind:

1. SJT questions look and act identically to Drag & Drop questions. But you must indicate that a Question is an SJT question if it is to be included and scored as such in a SJT Scenario Report. (See below)
2. You can flip a question type back and forth from SJT to Drag & Drop, when editing your Question sets.
3. Drag & Drop questions can have any number of responses. SJT questions must have 5 responses – this is a requirement of the scoring formulae.

We have found that some users are confused about what to do with the response bars, even if the instructions are clear. This is particularly so during an exam where they are sometimes too rushed to read the instructions.

Once you have seen how to drag the response bars into a different order with your mouse, it is fairly obvious. But the rectangular response bars do not always visually suggest what to do, for all users.

As a workaround to this, we have found that if you preface each response with a short HTML code, you can provide a visual hint about the drag/drop action needed. We have used the following HTML codes for this purpose:

* “&#8645; “ – this shows the ⇅ symbol.
* “&#8597; “ – this shows the ↕ symbol, but in some browsers, it shows a left arrow instead so this is less reliable.
* “&#9755; “ produces a right pointing finger. ☛

Remember to include the trailing semicolon in the above strings.

The good news is that if you are using a mixture of SCT and SJT and other questions in your exams, you can still use the same map[[1]](#footnote-1). It does not matter in which order the questions are posed within the exam map. SCT and SJT questions can be freely interspersed or even included on the same Nodes. So for the exam design, it is possible to have a common stem and context, describing a particular patient or situation, and then to ask SCT, SJT or other question styles, all on the same Node or page.

## How to set up a SJT Report

In a similar manner to how we use Scenarios in Script Concordance Testing (SCT), you must also use an OpenLabyrinth Scenario in order to produce a table of SJT scores. For more information on Scenarios generally, please consult XXX

Just as with SCT scoring, you will actually need to set up two Scenarios: one to store the scores of your expert panel or reference group; and one to store the scores of your exam participants. The good news is that if you are using a mixture of SCT and SJT and other questions in your exams, you can still use the same two Scenarios: one for the experts and one for the examinees. It does not matter in which order the questions are posed within the exam map. SCT and SJT questions can be freely interspersed or even included on the same Nodes.

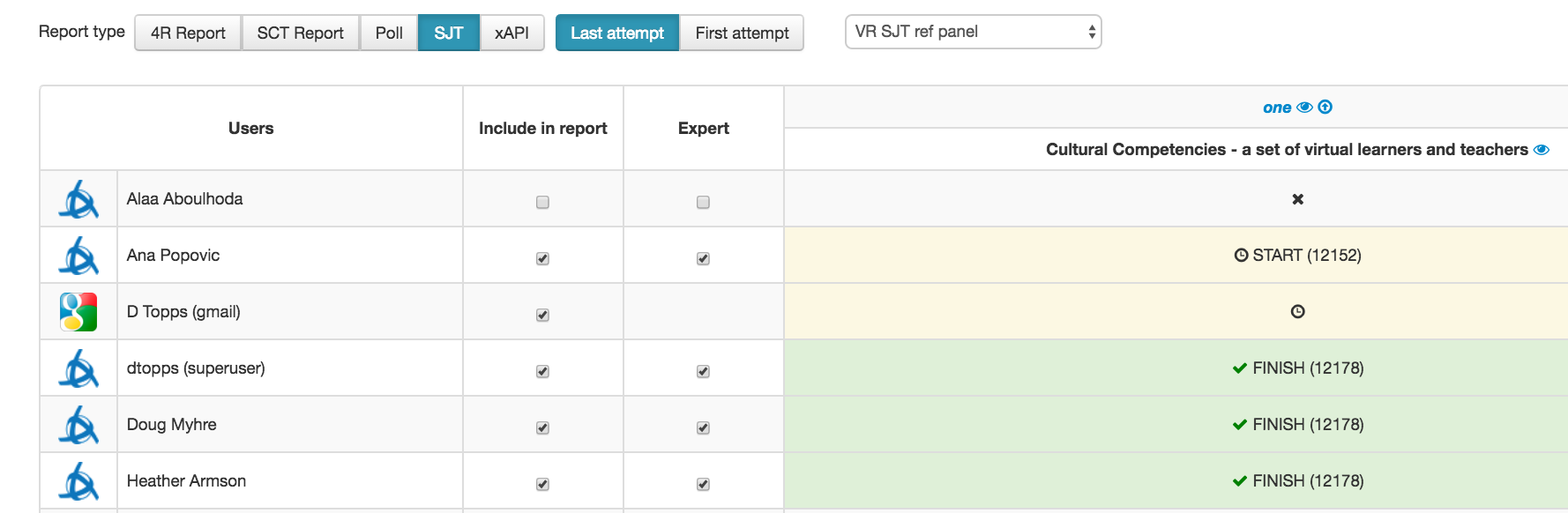
So for the exam design, it is possible to have a common stem and context, describing a particular patient or situation, and then to ask SCT, SJT or other question styles, all on the same Node or page.

A Scenario is simply a set of maps and a set of users, all in a single unit. For assigning sets of Users, you can make use of User Groups. This is a quick way to repeatedly assign the same Group of Users to different Scenarios.

A Scenario can also hold several maps, and then can be grouped in Steps. This allows the Scenario Director more control over when the users can access different parts of the Scenario. So you can, for example, create a first step with a few practice questions, but not open up the main part of the exam until the Scenario Director is ready to move everyone along to that next Step.

The Scenario Director can choose not to include all users when running a Report of the scores. Simply use the check boxes to show which users should be counted. If this Scenario is to be used as a reference panel, remember to also check at least one of those users as an Expert.

When running a report to generate the scores in either a SCT or SJT exam, open the Scenario and make sure that some users are selected. Select the Report Type, which is likely to be SJT for now.



Some exams will allow examinees to go back and revise their answers on previous Nodes. (See YYY for more on how to control this). Choose whether to score their first attempt or their last attempt at a question.

Finally, and most importantly, do not forget to choose your reference panel. This is the Scenario that holds the scores of your expert group. When generating the scoring panel for the reference group itself, choose the same Scenario as the one you are now reporting on. That is, you would make your first expert panel self-referential.

For those who are interested, you can actually compare multiple reference panels with each other to see how this affects the overall scoring, for either SCT or SJT exams. Indeed, as you ran multiple iterations of an exam, you might accumulate a larger set of experts, or eliminate some outliers.

Once you are ready to run your report on the selected candidates, some of which must have finished the map, as shown by a green FINISH check mark, click on the small blue eye beside the map name. This will generate the SJT (or SCT, if you chose that option) report.

Note that the math steps required are quite time consuming and it will often take 5-10 minutes to generate a report. Depending on the settings on your OpenLabyrinth server, this will sometimes exceed the timeout limits and the report will not complete properly.

Depending on the complexity of the exam and the number of cases or questions, we often found we had to limit the number of candidates being processed to about 20-25 at a time. This is a bit of a nuisance but you can fairly easily combine the generated spreadsheets to bring these multiple iterations together.

This is not an ideal situation. If any group is interested, we can look into optimizing the underlying code so as to make the generation of these reports more time and resource efficient. Please contact us.

## How good is the SJT approach?

In our exams and assessments, we have become increasingly keen on question formats which move away from single-best-answer (SBA). While SBA has the longest track record and is the favorite of psychometricians and item-response theorists as being the most reliable of formats, in real life we are often faced with several possible solutions to a problem from which we have to choose the least worst alternative.

In principle, SCT and SJT both use an approach where the correctness of your response is compared to a reference panel of experts or near-peers. This has the advantage that the correctness is determined not by the question author but by an evidence-informed comparison with what others of similar experience might choose.

This raises some interesting questions that are currently being debated in the assessment world. For example, in SCT, the shape of the curve showing the relative preferences of the expert panel matters a lot. In general terms, one is looking for good but not perfect concordance between the experts. After all, if there is perfect concordance and all the experts agree, then this is no different from a SBA style question: only one answer is correct.

And if the peakiness of the curve is too flat, then it matters little which choice is made, which indicates a non-discriminatory question. So what number indicates an acceptable degree of peakiness (kurtosis[[2]](#footnote-2)) or concordance? Remember that these are non-normal curves or Poisson distributions. Indeed, there is an advantage is questions are skewed to left or right; otherwise a canny participant can score more highly by simply picking the middle options.

More importantly, what should be done if there is a saddle-shaped curve or bimodal distribution? Some experts maintain that such a question should be rejected, as it indicates a poorly crafted or ambiguous question. Others hold that this is what life is like and that such situations also arise in real life. This is keenly debated at present. It is up to the exam administrators as to what tack they wish to take in such circumstances.

Now we find something that is a little strange in how SJT scores are compared with the reference panel, especially since the two formats share similar basic principles. The responses of the reference panel still determine which is the most to least popular option. But no attention is paid to the concordance of the distribution. The way that the calculations work, the relative order of the top two choices is just as important as the bottom two choices.

But when you observe respondents making their decisions, are the bottom choices really as important? We find that for the majority of questions, a lot of attention is paid to the top two choices, less to the next comparison, with little regard for the last two choices, which are almost discarded as irrelevant or unworthy of consideration.

This is not always true. We sometimes observe what we call ‘negatively skewed questions’. This is where the choices are all relatively undesirable but some are worse than others and the respondent has to make the best of a bad lot. This type of question may indeed have value. There are situations where one needs to look for outliers who do not follow social or ethical norms. Sometimes, we are not selecting for the best candidates, but instead selecting against the marginal ones.

Another recommendation that we find odd is that SJT questions are supposed to be phrased such that the respondent is asked what they ‘should’ do, not what they ‘would’ do? Some groups who work in this area have assured us that they have data to support this approach. However, this somewhat goes against the approach suggested in ‘behavioural descriptive interviewing’ (BDI). In BDI, candidates are asked not what they should do in a given situation but to give an example of a similar situation and what they actually did. (Taylor & Small, 2002)

Now that we have much more flexible user interfaces at our disposal, it may be time to start asking for responses in more complex ways such as biaxial responses: asking the user to evaluate an option on more than one scale eg. Desirability vs Practicality. We have created an example map illustrating some of these possible approaches here: [‘SJT Concordance’ case](http://demo.openlabyrinth.ca/renderLabyrinth/index/539). (Use ‘SJTconc’ when asked for a key.)

There are many aspects of SJT that remain worthy of further exploration. We are keen to explore these potential variations with groups who would like to use OpenLabyrinth as an SJT platform.

## References

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O’Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education*, *46*(9), 850–868. http://doi.org/10.1111/j.1365-2923.2012.04336.x

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology*, *75*(3), 277–294. http://doi.org/10.1348/096317902320369712

1. *The terminology that we use in OpenLabyrinth for maps and cases can get confusing here. Most of the time, a map (its correct term) describes a single virtual patient case, with many branching choices for that patient. And so, it is often also referred to as a case. However, in SCT and SJT exams, a case is more likely to refer to single patient stem, which may have 2-4 questions arising from it. In this event, a case is more likely to be confined to a single Node.* [↑](#footnote-ref-1)
2. Kurtosis scores themselves are hard to follow, not aligning with usual variation from 0 to 1. In true terms, they are only an approximation of peakiness. [↑](#footnote-ref-2)