

Research Summary: PTE Academic Sensitivity Review Project

1. Introduction

The main aim of the Pearson Test of English Academic (PTE Academic) Item Sensitivity Project was to review the contents of the initial item bank of PTE Academic with a view to detecting and remedying any instances of predictable bias against, or in favor of, particular groups of test takers. Items deemed inappropriate could then either be edited and changed or removed from the item bank. A subsidiary aim was to establish a methodology for conducting such reviews for future items.

2. Background

To ensure impartiality, Pearson contracted an external expert, Professor Fred Davidson of the University of Illinois at Urbana-Champaign, to chair a panel of reviewers. After consultation with Professor Davidson, 15 reviewers from 14 different countries were recruited:

Brazil	Japan
China	Korea
Costa Rica	Morocco
Finland	Slovenia
Hungary	Taiwan
Indonesia (two representatives)	Ukraine
Israel	United Arab Emirates

The countries were chosen so as to achieve a spread of regions and cultures. The two panelists from Indonesia in fact shared the workload equally and so constitute the equivalent of a single reviewer. The reviewers were all highly proficient in English and had worked for many years in their respective countries as English language teachers and, in some cases, as test developers as well.

A set of reviewing guidelines were developed for the project and sent out to the panelists. This included the following statement of the overall goal of the project:

'The Bias-Sensitivity Review Panel needs to make recommendations concerning sensitivity to different cultures, religions, ethnic and socio-economic groups, and disabilities, gender roles, use of positive language, symbols, words, phrases and content, and whether an item requires field-specific knowledge. In general, the review is to detect items and text that introduce construct-irrelevant variance, or elicit strong emotional response by members of racial, ethnic, gender, or other groups, or a strong reaction due to personal factors and, as a result, may prevent those groups of test takers from accurately demonstrating their English skills.'

3. Item Review Process

The items were sent to the panelists who were asked to rate them according to a three-point scale as follows:

0. the item is not sensitive; it may be kept in its present form
1. the item is sensitive but can easily be edited so as to remove the sensitivity
2. the item is sensitive and cannot easily be edited to remove the sensitivity

Generally, an item was expected to be rated 1 where the source of sensitivity was localized within the editable text of the item, such that it could be removed by deleting or substituting a few words. An item would be rated 2 if the source of the sensitivity could not be easily edited out. For example, the sensitivity might be distributed throughout the item, or it might be found in an audio or video recording. It was expected that editing audio or video material whilst maintaining the integrity of the item, would be difficult.

The items were distributed among the 15 panelists so that each item was reviewed by two panelists, one from the Eastern and the other from the Western Hemisphere. The panelists received the text of the items together with any reading texts, audio and video files, audio and video transcripts and graphics pertaining to the content of the items.

Each panelist reviewed their allocated items independently and logged their rating — 0, 1 or 2 — for each item, together with their comments in the case of items which they rated 1 or 2. The ratings and comments were returned to Professor Davidson in his capacity as Chair of the review panel.

All items that were rated 0 by both reviewers were deemed to be unproblematic with regard to sensitivity and so were not subjected to further scrutiny. This accounted for 83.5% of the items reviewed. In all other cases, that is the 16.5% of items which at least one reviewer had rated 1 or 2, the Chair adjudicated the reviewers' decisions. For each of these items the Chair made one of three recommendations:

- keep: the item is not sensitive and can be kept without change
- edit: the item is sensitive and should be edited so as to remove the sensitivity
- drop: the item is sensitive and cannot be edited, so should be removed from the item bank

4. Results

Table 1 Results of panelists' ratings and the Chair's adjudications.

Sumrec	Percent	Cumulative percent
1edit	15.49	15.49
1keep	37.09	52.58
1drop	8.45	61.03
2edit	3.05	64.08
2keep	13.38	77.46
2drop	15.26	92.72
3edit	0.47	93.19
3keep	1.64	94.84
3drop	3.52	98.36
4keep	0.23	98.59
4drop	1.41	100.00

In this table the variable 'sumrec' is a composite of the sum of the panelists' ratings and the Chair's recommendation. For example '1edit' represents those which only one panelist rated as 1 and which the Chair recommended to be edited; '2keep' represents those items which were given a total rating of 2 (either each panelist rated the item as 1 or one rated it as 2 and the other as 0) and the Chair's recommendation was to keep.

It can be seen that the most common result is '1keep', meaning that one of the two panelists found some editable sensitivity in the item but the Chair judged that the item could be kept without editing. In only a very small number of cases (1.41% of adjudicated items) was there a unanimous recommendation that the item should be removed (4drop).

5. Actions

Following Professor Davidson's recommendations, those few items where there was a unanimous 'drop' recommendation ('4drop' in Table 1) were consequently removed from the item bank.

Those items that Professor Davidson judged not to be sensitive (i.e., all those tagged with the recommendation 'keep'), were convincingly supported by Davidson's arguments and the items were left in the item bank.

The remaining items were provisionally withdrawn from the item bank and subjected to a statistical review process conducted by Dr Joshua Goodman of James Madison University, Virginia, USA. The aim of this phase of the research was to use data from the PTE Academic field tests to ascertain, where possible, whether and to what extent test takers' performance had in fact been affected by item sensitivity.

The items were first classified according to which group or groups of test takers were likely to be affected by the type of sensitivity which had been flagged in the review process. Where the group classification corresponded to demographic information (such as age, gender or nationality) which had been collected from test takers in the course of field testing, and where there was considered to be sufficient such data to make statistically relevant comparisons, the items were earmarked for statistical analysis. This left a substantial number of items which were not considered to be amenable to statistical review for one of three reasons:

- Sensitivity was not related to any particular grouping. These items were said to display 'general sensitivity'.
- The group identified was one about which no information was available in the field test demographic data. This applied to, for example, 'children from broken homes' or 'animal rights activists'.
- There was insufficient field test data available to form statistical groups. This applied to certain nationalities that were not strongly represented in the field test data.
- All items which were not considered to be amenable to statistical review were subjected to a final internal review by PLT staff. This is described below.

6. Statistical Review

Statistical review consists of computing DIF (Differential Item Function) characteristics in relation to the grouping variable that has been identified for each item affected. This consists of estimating the item difficulty separately for the members and non members of the group of subjects that would supposedly be affected by the issue flagged as sensitive. For example, if an item is flagged as possibly being sensitive for female students, the whole response data set is divided into responses from female subjects and those from male students. Subsequently, the item difficulties of all items are estimated separately within these two groups. If items are equally fair for both female and male subjects, the item difficulty estimates are expected to be the same but for measurement error. If the difference between the two separate difficulty estimates is larger than twice the standard error of estimation (SEE), the item is considered to be statistically biased.

The outcome of statistical analysis was that only four items met this criterion. These items were then dropped from the item bank.

7. Final Internal Review

The remaining items (those deemed not to be amenable to statistical analysis) were scrutinized by two expert members of the Test Development team. This scrutiny was undertaken independently by the two staff members, without conferring with each other. The outcome for each item was a recommendation to return it to the item bank ('keep') or remove it permanently ('drop'). Where the two recommendations coincided, the corresponding action was taken and the item was kept or dropped. Where the recommendations conflicted, a third member of staff delivered a final adjudication.

As a result of this final review, 37% of the internally reviewed items were deemed to be sensitive and if the cause of the sensitivity was considered to be irreparable, they were dropped from the item bank. The items were considered irreparable if the sensitivity issue was pervasive, or if removing the sensitivity would have required editing written or audio prompts beyond the limits of authenticity. A further 13% of the reviewed items were judged to be editable, that is to show sensitivity that could be removed by minor edits to the item text. These items will be removed from the item bank, edited, then re-introduced as new items (thus they will be field tested before they can be used in live tests).

8. Conclusion

The PTE Academic (PTE A) Sensitivity Review Project revealed that a small proportion of test items were considered to be potentially biased against particular groups of test takers. In turn, for a very limited number of those items, statistical evidence of such bias was indeed found. In the course of the review process a set of reviewing guidelines was developed and validated that will be used to review future items. In order to reduce the incidence of item sensitivity in the future, the lessons learned in the review process will be incorporated into test specifications and will be applied in the training of item writers and reviewers.

Appendix 1:

Sensitivity Review end-to-end process

