
ALWAYS
LEARNING

Avoiding Construct-Irrelevant Variance

Or the need for sensitivity review guidelines during item development

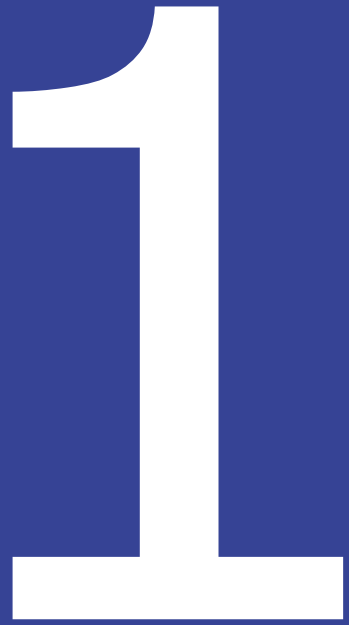
Kirsten Ackermann

EALTA 2011, Siena

Outline

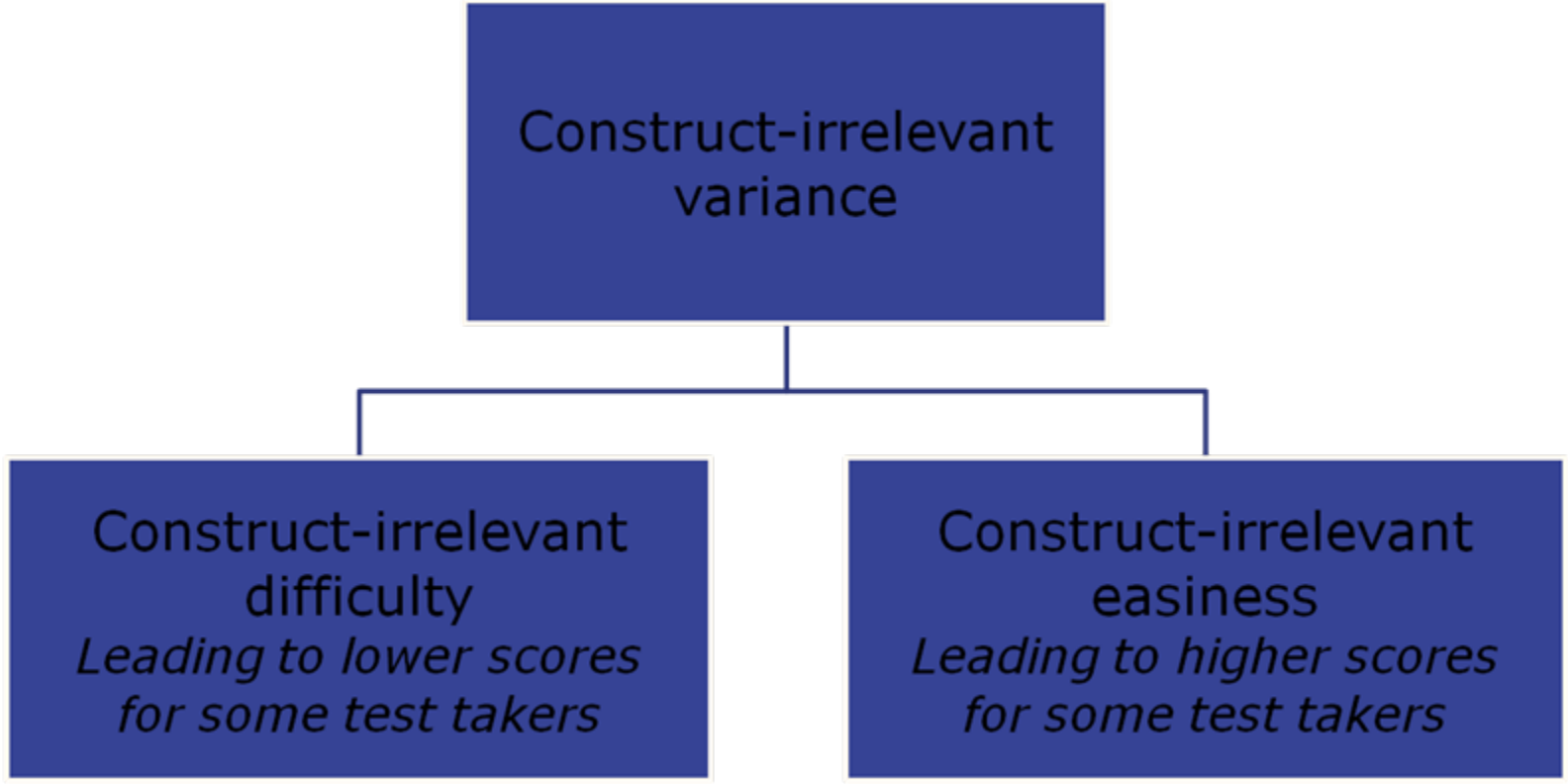
1. Background
2. Sensitivity Review Project
3. Sensitivity Review in Practice

Background



Messick (1989) on Construct-irrelevant Variance

Construct-irrelevant variance exists when the test is measuring something that is irrelevant to the theoretical construct that the test is supposed to be measuring.



Motivation

The target test taker population of PTE Academic is heterogeneous - speaks different languages, comes from various cultural and social backgrounds, and studies or intends to study a wide variety of academic subjects.

To detect and remedy any instances of predictable bias in an existing item bank which can result from e.g., cultural, religious or gender differences among the test population and may prevent particular test taker groups from accurately demonstrating their language skills.

To develop sensitivity review guidelines and implement a standard review process.

Sensitivity Review Project

2

PEARSON

Phase One: Panel Review
The Panel

Chair

- Fred Davidson, Professor, Dept. of Linguistics, University of Illinois at Urbana-Champaign

Panelists

- 15 people representing 14 distinct nations and regions:

  Hungary, China, Brazil, Ukraine, Japan, Morocco, Korea, Indonesia, Slovenia, Israel, United Arab Emirates, Taiwan, Costa Rica, Finland

- Spoke the national language of the country they represented as their mother tongue
- highly proficient in English
- Extensive experience in teaching ESOL and in some cases as test developers

Phase One: Panel Review Methodology

Objective:

"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 a 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."

Phase One: Panel Review
Methodology

Review guidelines:

- Formulating review guidelines (Chair and Pearson staff)
- Guidelines were informed by:
  - ETS Standards for Quality and Fairness (2002)
  - ETS Fairness Review Guidelines (2003)
  - ETS International Principles for Fairness Review of Assessments (2007)
- Reviewing and revising guidelines
- Developing a 3-point rating scale:
  - 0 = no sensitivity
  - 1 = sensitive, but item can be altered to remove sensitivity
  - 2 = sensitive, but item cannot be easily altered to remove sensitivity

Recommendation to remove item from item bank

Phase One: Panel Review Methodology

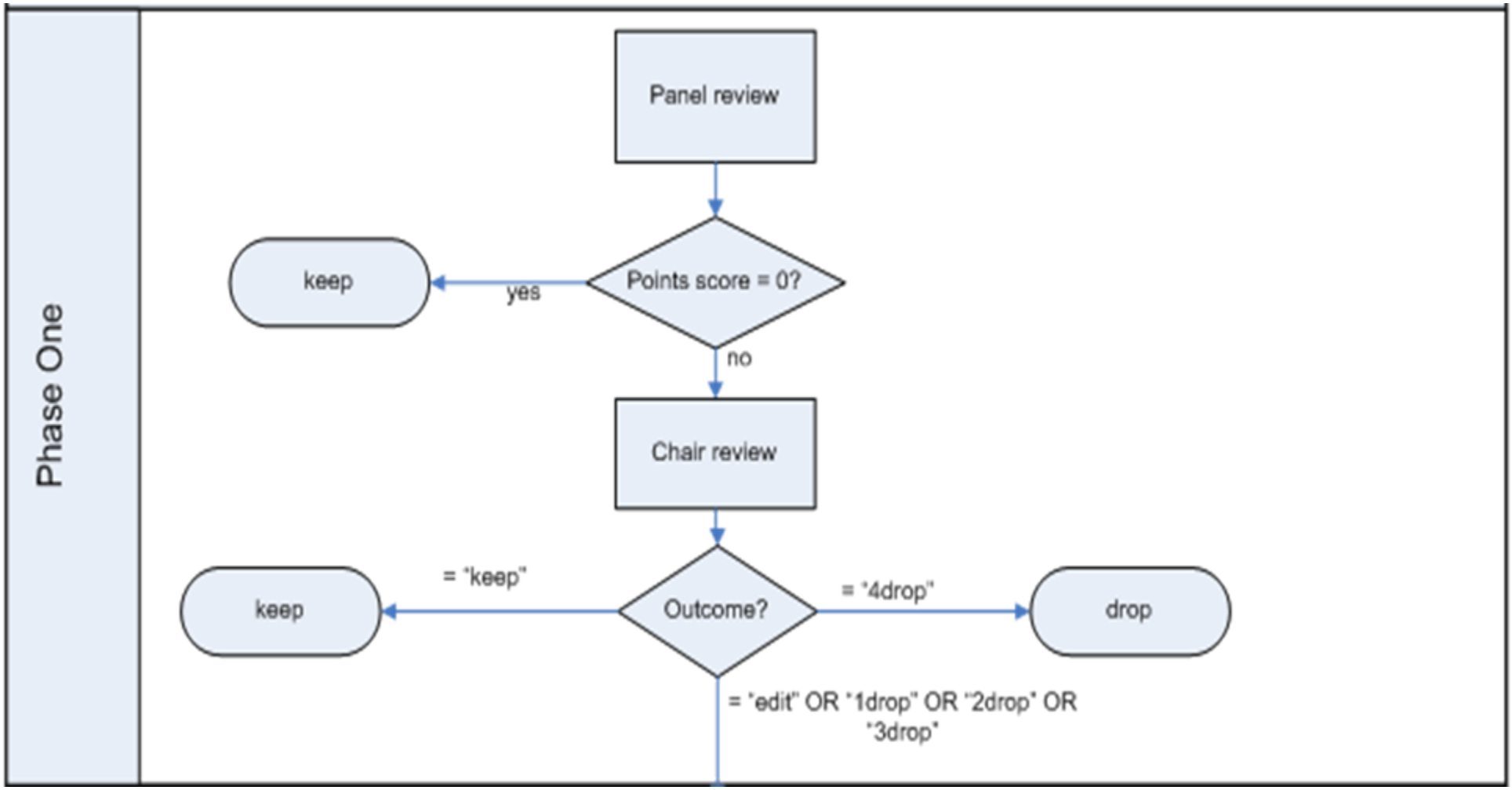
Reviewing process:

- Panelists received items including directive, passage/transcript/prompt, options, and where necessary, audio/video and graphic item material

  ➡ Audio and graphic material allowed panelists to judge sensitivity related to the mode of delivery and to avoid over-reliance on transcripts.

- Each item was reviewed independently by two panelists and adjudicated by the chair where at least one panelist rated other than ‘0’.

Phase One: Panel Review Methodology



PEARSON

Phase One: Panel Review
Results

Recommendations	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

Adjudicated items: Summation of panelists ratings followed by Chair's recommendation

Phase One: Panel Review

Results

Comparison between the practice of journalists and scientists

Recommendations	Percent	Cumulative percent
1edit	15.49	15.49
1keep	37.09	52.58
1drop		3
2edit		3
2keep	13.38	77.46
2drop	15.26	92.72
3edit		.19
3keep		.84
3drop		.36
4keep	0.23	98.59
4drop	1.41	100.00

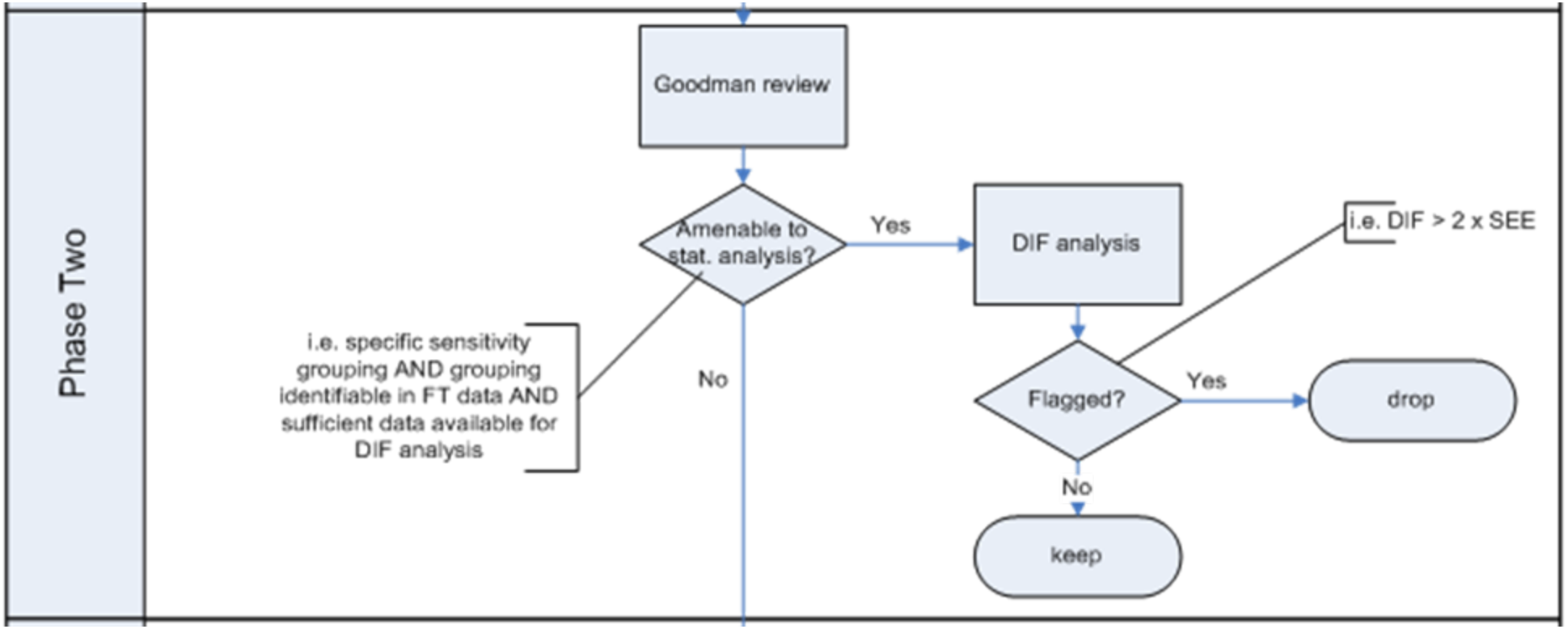
Insect-eating diet
Financial practice (debts and loans)

Ethnic cleansing
Violence
Cardiac depression

Adjudicated items: Summation of panelists ratings followed by Chair's recommendation

Phase Two: Statistical Analysis

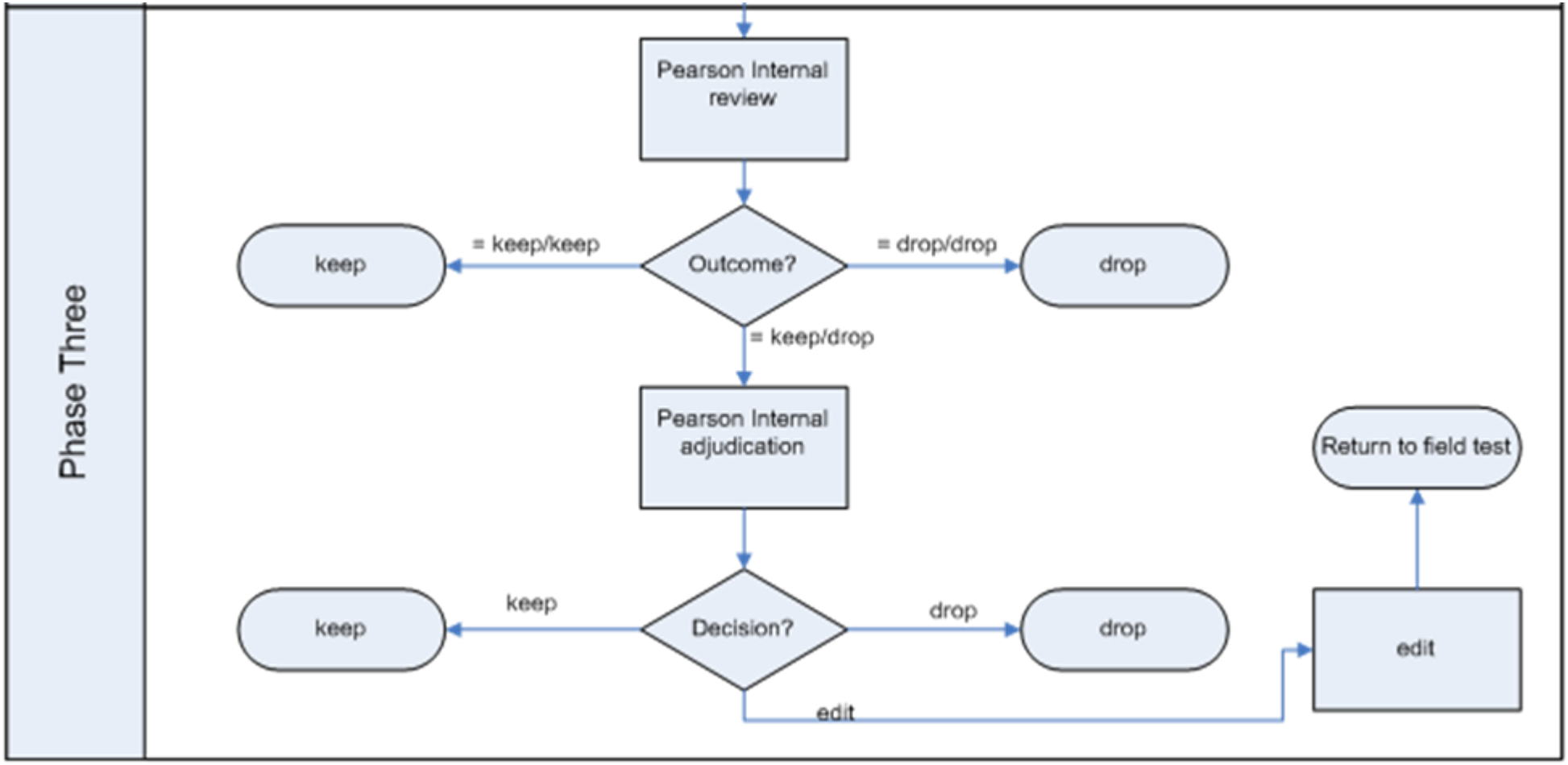
Items that were recommended for editing underwent statistical analysis conducted by Dr Joshua Goodman of James Madison University, Virginia, USA



PEARSON

Phase Three: Internal Review

All items which were not considered to be amenable to statistical review were subjected to a final internal review.



PEARSON

Sensitivity review in practice

3

PEARSON

Practical implications: Sensitivity review guidelines

Sample 10
Item Keywords: 16-LS-REPT student advisors
Accession: I02761.2
Type: Fill the blank

*DIRECTIVE*You will hear a sentence. Please repeat the sentence exactly as you hear it. You will hear the sentence only once.

*TRANSCRIPT**Female* mentors are extremely helpful.

Analysis Sample 10

This is a **focused sensitivity**, derived from both the **gender** reference in the content and the **style and rhetoric** of the speaker. The sentence, which is meant to be repeated by the test taker, is pronounced with such **irony** that the test taker must believe that female mentors are of no help at all. This clearly reveals that the speaker has gender related prejudices and the item must therefore be flagged as sensitive. Since correction would mean the re-recording of the sentence, this item should be dropped from the item bank.

Practical implications: Item development workflow

- Sensitivity review guidelines were adapted for PTE General.
- Item writers and reviewers are trained in applying the sensitivity review guidelines when writing or reviewing items.

The item is written by an item writer who may be in the United Kingdom, the United States, or Australia.

The item is reviewed by another item writer in a different country.

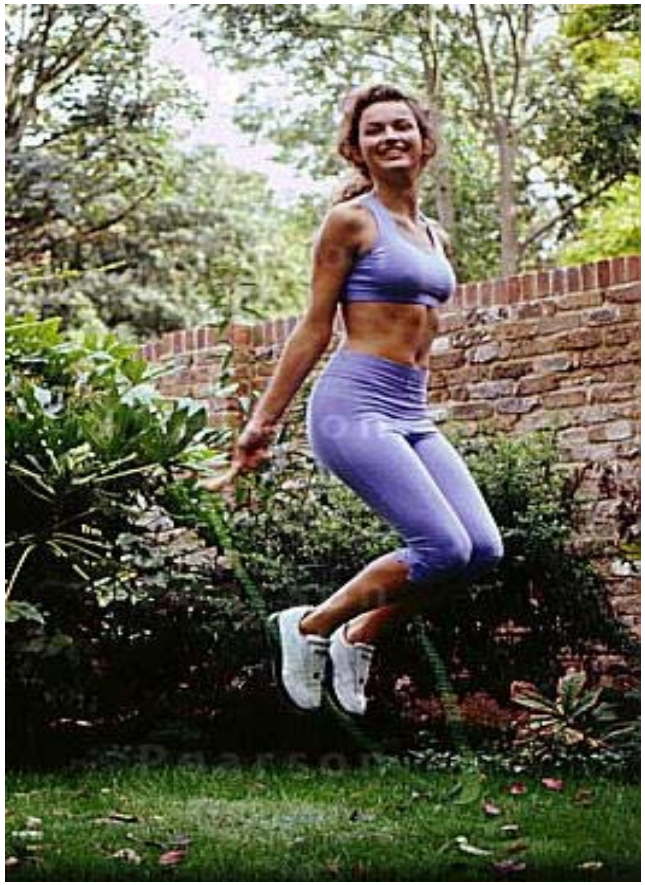
The item is reviewed by an expert team who are not involved in the item writing to ensure that it meets the requirements of the test specification.

If the item is flagged in the previous stage, it is reviewed by the PLT Test Development team to ensure that it meets the requirements of the test specification.

The PLT test development team review the item to ensure spelling and grammar are correct.

PEARSON

Samples of sensitive items
Speaking task: Picture description



Promotion of healthy living

Listening task: SAMC

PROMPT Listen to the conversation. What are the speakers doing?

V1(f): I can't believe some of the things people wear today! It's so different from when we were growing up.

V2(m): I know. I was out in town last night. It's the middle of winter and there were girls wearing short skirts and nothing on their legs.

V1(f): And no coats either, I imagine. It seems to be the fashion these days…

Writing task: Write text

PROMPT Write an essay on the following topic.

There are limits to scientific knowledge because some things can never be explained by scientists.

How far do you agree with this statement? Write 250-300 words

Results
Rejection rates

PTE Academic Commission	Dropped d/t sensitivity
2009	7.4%
2010	2.0%

PTE General Commission	Dropped d/t sensitivity
2010	2.4%
2010	1.7%

- Compiling a comprehensive report on item writing session
- Providing country reports
- Providing feedback to individual item writers
- Highlighting areas for additional training

Thank you

kirsten.ackermann@pearson.com