Appropriate Item Analysis for Continuous Data

Prof Clarke Hazlett

Chinese University of Hong Kong

Hong Kong SAR, China

Overview of Presentation

Part I

Introduction: What is Item Analysis (IA)?

Part II

Review of IA for items marked as right=1 or wrong=0

Part II

IA for items marked in a continuous manner (e.g., from 0 to any value such 99.9)

How to Interpret & Use Results to Improve Items

Introduction: What is Item Analysis?

Part I

Item Analysis (IA)

Process by which assessment items are critically reviewed

- determine if items function according to expectation
- identify structural flaws
- improve item quality

Using both expert judgment & empirical data

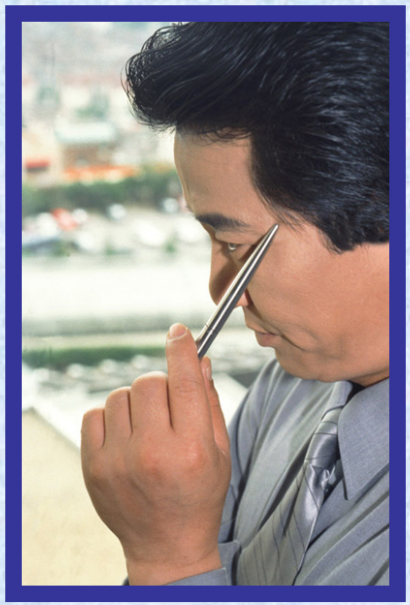
Judgmental Item Analysis

Expert Judgment addresses these queries

Are content, processes & constructs being assessed by the item relevant?

Is the item properly structured?

Is the item free of bias?



Empirical Item Analysis

Relevant psychometric properties are empirical data

Item's difficulty level

Item's correlation with the total mark on the assessment

(or correlation between item & a reference/gold standard)

Item's ability to discriminate between poorer & better students

Distracter analysis: trends in how students answered the item

Review: IA for Items Scored as Right/Wrong

Part II

Student Responses Marked in a Binary Manner (0/1)

Most selected response formats (MCQs) are marked as right (1) or wrong (0)

X - type (True / False)

Multiple X - type (Multiple T / F)

A - type (best one of n options)

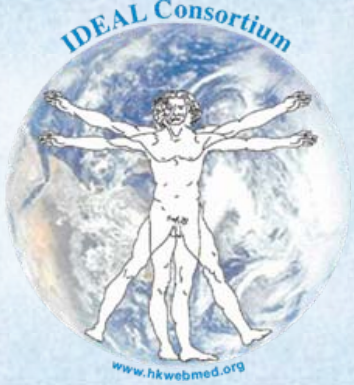
R - type (extended matching)

Typical Statistical Report for an Item Scored as Right (1) or Wrong (0)

ITEM 1: DIF=0.837 RPB= 0.179 CRPB= 0.049 95% CON = [- 0.125, 0.220]

GROUP	N	A	B *	C	D	E
TOTAL	129	.12	.84	.00	.01	.04
HIGH	39	.05	.95	.00	.00	.00
MID	58	.12	.81	.00	.02	.05
LOW	32	.19	.75	.00	.00	.06
DISCRIMINATING POWER		- 0.14	0.20	0.00	0.00	- 0.06

* correct answer



How does one use these statistical reports

to flag & diagnose potentially flawed items?

Example how IA Identifies Item Writing Flaws: 1st Version Item

Among the common study designs used in clinical research, a study of assumed harmful effects of an intervention requires use of which design in order to establish the most valid but also ethically obtained evidence?

A. case study
B. case series
C. case-control (retrospective) study *
D. cohort (prospective) study
E. randomized controlled trial

Statistical Report for the Original Item

ITEM 30:	DIF=0.38	RPB= 0.296	CRPB= 0.139	95% CON = [- 0.035, 0.304]

GROUP	N	A	B	C *	D	E
TOTAL	129	.16	.00	.38	.47	.00
HIGH	39	.05	.00	.53	.42	.00
MID	58	.22	.00	.41	.37	.00
LOW	32	.16	.00	.16	.69	.00
DISCRIMINATING POWER		- 0.10	0.00	0.37	- 0.27	0.00

Revised 2nd Version of Item

Among the common study designs used in clinical research, a study of rare, assumed harmful effects of an intervention requires use of which design in order to establish the most valid but also ethically obtained evidence?

A. case study
B. case series
C. case-control (retrospective) study*
D. cohort (prospective) study
E. randomized control trial (RCT)

Statistical Report for Revised Item

ITEM 1: DIF=0.837 RPB= 0.179 CRPB= 0.049 95% CON = [- 0.125, 0.220]

GROUP	N	A	B*	C	D	E
TOTAL	129	.12	.00	.84	.01	.04
HIGH	39	.05	.00	.95	.00	.00
MID	58	.12	.00	.81	.02	.05
LOW	32	.19	.00	.75	.00	.06
DISCRIMINATING POWER		- 0.14	0.0	0.20	0.00	- 0.06

IA for Items Marked as Continuous Data

Part III

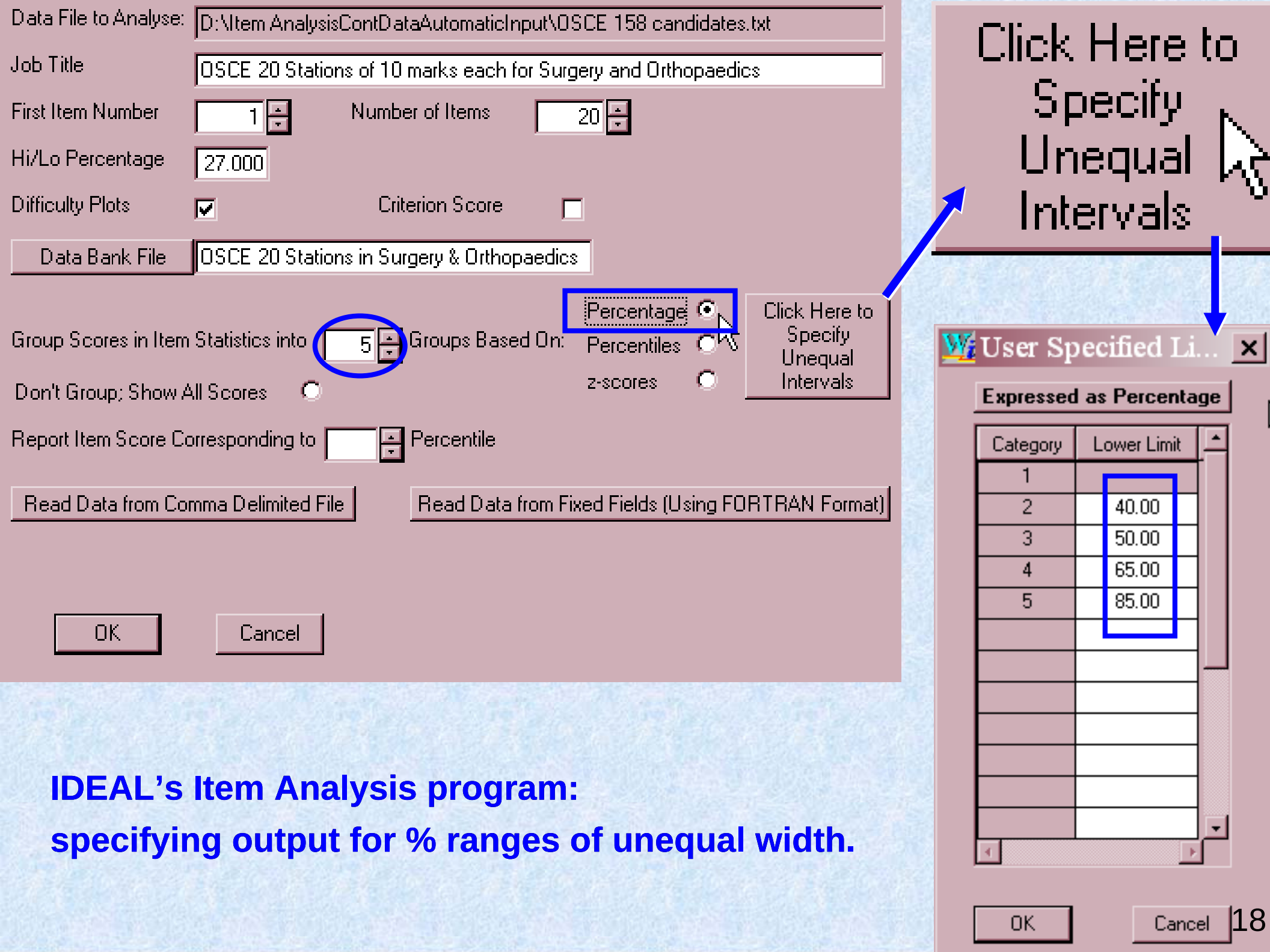
Items Marked as Continuous Data: 0 to any value

Most Constructed Response Formats

Modified Essay Questions	(MEQ)
Short Answer Questions	(SAQ)
Objective Structured Clinical Examinations	(OSCE)
Objective Structured Practical Examinations	(OSPE)
Orals	(Viva)
Class & Poster presentations	(Projects)
Research	(Reports)

IA Discrimination Matrices for Continuous Data

- **Instead of options (as used with IA for MCQ items), performance categories (reflecting percentage ranges) are used to reveal trends in how students performed**
- **These ranges in performance can be**
  - equivalent or non equivalent in width
  - and are more useful if the widths match cut points for clear fails, borderline fails, borderline passes, clear passes & clear distinctions



IDEAL's Item Analysis program:

specifying output for % ranges of unequal width.

IA Example for OSCE Stations

Example: Station 4 in a 20 Station Surgery OSCE in 5th Yr MBChB

Counseling patient management: skin lesion needing excision

AT THIS STATION: PHOTOGRAPH OF A PATIENT'S CHEEK WITH A LESION & A FAX SENT BY THE PATIENT'S RELATIVE. READ THE PATIENT'S DETAILS BELOW & MAKE PHONE CALL TO SON-IN-LAW ACCORDING TO INSTRUCTIONS.

History:

Mrs Wong, 70 yr-old, has come to Outpatient Clinic because of a growth on her cheek. She lives alone, is rather forgetful, & so has asked you to ring her son-in-law who lives in Singapore. His fax has details of what he feels he needs to know so that the family can advise Mrs Wong appropriately.

Telephone conversation **[10 marks]**

Station 4: Scoring Instructions for the Examiner

1. Introduction: 1 mark [details not on this slide]

Candidate introduces him/herself & clarifies s/he is looking after Mrs Wong.

2. Is this a cancer? 2 marks [scoring details below]

Mrs Wong has a typical seborrheic keratotic lesion which is benign; common among old people. Morphologically are neoplasms with variable melanin pigmentation.

Score 2: correct diagnosis & conclusion all expressed in lay language
Score 1: reasonable alternative explanation but conveys same message
Score .5: misleading answer given inaccuracies and poor explanation
Score 0: meaningless information & poor communication

3. Does it have to be removed? 1 mark [scoring details . . .]

4. What would happen if not removed? 1 mark [scoring details . . .]

5. Would she need to be hospitalized? 1 mark [scoring details . . .]

6. Would there be a scar? 1 mark [scoring details . . .]

7. Inquire about the patient's use of aspirin? 1 mark [scoring details . . .]

IA Report for Station 4 Using Unequal % Ranges

% Range: unequal ranges of performances base on a school's cut points

ITEM 5: DIF=0.798 CORR= 0.404 CR_R= 0.214 95%CON = [0.058, 0.359]

GROUP	N	performance range: < 40%	40-49%	50-64%	65-84%	> 84%	Overall
TOTAL	156	0.01	0.03	0.16	0.26	0.53	0.80
HIGH	45	0.00	0.00	0.04	0.22	0.73	0.87
MID	69	0.01	0.01	0.13	0.33	0.51	0.81
LOW	42	0.02	0.10	0.33	0.19	0.36	0.70
DISCRIMINATING POWER:		- 0.02	- 0.10	- 0.29	0.03	0.38	0.17

Station 4: Counseling patient management: skin lesion needing excision

IA for Station 5 Using Unequal % Ranges (cont'd)

Example: Chosen Unequal Percentage Ranges

% Range	< 40%	40-49%	50-64%	65-84%	> 84%	
	1	2	3	4	5	Overall
TOT	0.01	0.03	0.16	0.26	0.53	0.80
HI	0.00	0.00	0.04	0.22	0.73	0.87
MID	0.01	0.01	0.13	0.33	0.51	0.81
LOW	0.02	0.10	0.33	0.19	0.36	0.70

Station 4: Counseling patient management: skin lesion needing excision

Summary of Interpreting IA for Station 4

- Mean, Correlation & overall Discrimination Power indicate station was okay
- Discrimination matrix indicates station has discriminated well and from an educational viewpoint, skill has been adequately taught and/or learned
  - Only 1% are clear failures (these were in low group)
  - Another 3% are borderline failures (almost all were in low group)
  - All high & most in middle groups were above the cut point for passing
- This station measures what the overall OSCE measures (i.e., clinical skill)

Station 5: Breaking Bad News: family of terminal cancer patient

YOUR PATIENT'S RELATIVE IS AT THIS STATION. PATIENT UNDERWENT AN EXPLORATORY LAPAROTOMY AT WHICH AN INOPERABLE CANCER OF THE STOMACH WAS CONFIRMED BY BIOPSIES TAKEN FROM LIVER & STOMACH.

RELATIVE IS PATIENT'S SON WHO WISHES TO KNOW NOW WHAT TO EXPECT. RELATIVE WILL OPEN THE CONVERSATION**.**

History

Patient is 58 yr-old male referred by his doctor to Surgical Outpatient Clinic with epigastric pain, anorexia, considerable weight loss & general weakness; admitted soon after for investigation & eventual laparotomy.

As a result of these investigations the family already has been told that the tumor was advanced & there would be nothing gained from an operation; however, the family insisted that patient should be given a chance. Unfortunately, results of the past investigations were confirmed at operation.

Scoring Instructions: Maximum score: 10

8/10: Excellent; 6/10: Pass; 4/10: Inadequate or weak

IA Report for Second Communication Station

Difficulty = .658 Corrected correlation = .084,

% Range	< 40%	40-49%	50-64%	65-85%	> 84%	Overall
TOT	0.00	0.12	0.47	0.41	0.00	0.66
HI	0.00	0.09	0.36	0.56	0.00	0.69
MID	0.00	0.12	0.48	0.41	0.00	0.66
LOW	0.00	0.17	0.57	0.26	0.00	0.62
Discrim	- .00	- .08	- .21	.30	.00	.07

Station 5: Breaking bad news: terminal cancer

Interpreting IA for Station: Breaking Bad News: Cancer

- Mean (66%) is acceptable & above cut point for passing (desirable)
- Corrected correlation (.08) is not different from zero (indicating this communication station measures something different than overall OSCE)
- Discrimination is only 0.07 (not adequate)
  - 9% of OSCE's top performers fail & none are outstanding in this station)
- Why is this communication station so different than previous communication station?
    - Is station mismatched to skill level of students and/or the teaching?
    - Does station fail to adequately simulate communicating bad news?
    - Is there a language problem?
    - Were scoring instructions inadequate?

Interpreting IA for Station: Breaking Bad News: Cancer

We determined that the scoring system provided for the markers was problematic

- Categories for assigned marks were too broad
  Scoring Instructions: 8/10 – Excellent; 6/10 – Pass; 4/10 – Inadequate
- Used nurses as markers in this station; previous stations used doctors; nurses were apparently very reluctant to assign scores over a broad range

Final Comment on Item Analyses

Begin with the judgmental method & ensure responses to following questions are positive:

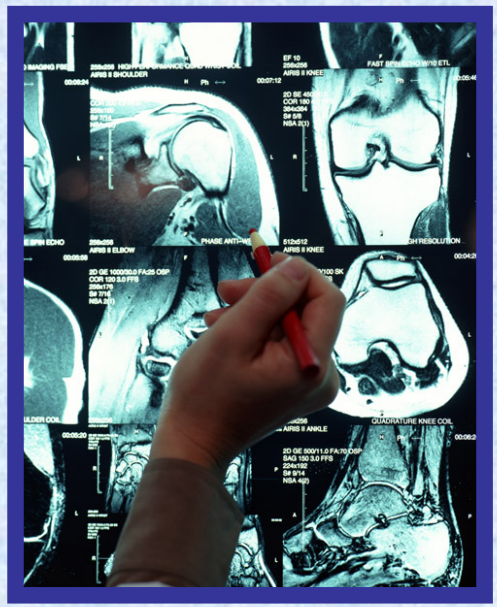
Is content and process contained in the item relevant?

Is item properly structured?

Is item free of bias?

If response to any question is negative,

take corrective measures.



Then Consider the Item's Psychometric Properties

Is item of appropriate difficulty?

Is item – total test score correlation positive?

Is discrimination power positive for the best answer in MCQs & in the high performance ranges if using essays, OSCEs, short answer questions, etc?

Is discriminating power negative for each distracter in MCQs & in the low performance ranges if using essays, OSCES, short answer questions, etc?

Are these performance characteristic consistent with the purpose of your assessment?

References:

Case S.M. & Swanson D.B. (2001).

Constructing written test questions for the basic and clinical sciences. Philadelphia: National Board of Medical Examiners**.**

Osterlind S.J. (1998).

Constructing test items. Boston: Kluwer Academic Publishers.

Precht, D., Hazlett, C., Yip, S. & Nicholls, J. (2005)

IDEAL – HK™ Item Analysis Users' Guide: Selected and Constructed Item Formats. Hong Kong: Candor Production Ltd