Putting a G-theory approach to marking reliability through its paces

Ben Smith

Background

- Increasing focus on marking quality in English assessments
  - High stakes for both candidates and schools (and increasingly so)
  - Consequences of grade misclassification can therefore be serious

- In GCSE and A-level assessments (sat at 16 & 18 respectively) live marking is mark-remark monitored
  - Seeding = senior examiner sets a mark for a response, ‘seeds’ are secretly interspersed into other examiners’ normal marking
  - (Sample) double-marking = a proportion of responses (usually 5%) are marked by two examiners
  - Post hoc, classical statistics can be computed from this data

Example of seeding data

Candidate responses (seeds)	markers								
	1	2	3	4	5	6	7	8	9
1	x	x			x	x	x		x
2	x	x	x	x	x	x	x	x	x
3	x	x	x	x		x	x		
4					x			x	
5	x	x	x	x	x	x	x	x	x
6			x				x		
7	x	x	x	x	x	x	x	x	x
8			x	x	x			x	
9		x	x			x	x	x	
10		x		x	x		x		x

Example of double-marking data

Candidate responses	markers								
	1	2	3	4	5	6	7	8	9
1	x						x		
2					x				x
3			x			x			
4					x			x	
5	x				x				
6			x				x		
7				x		x			
8			x		x				
9			x					x	
10				x	x				

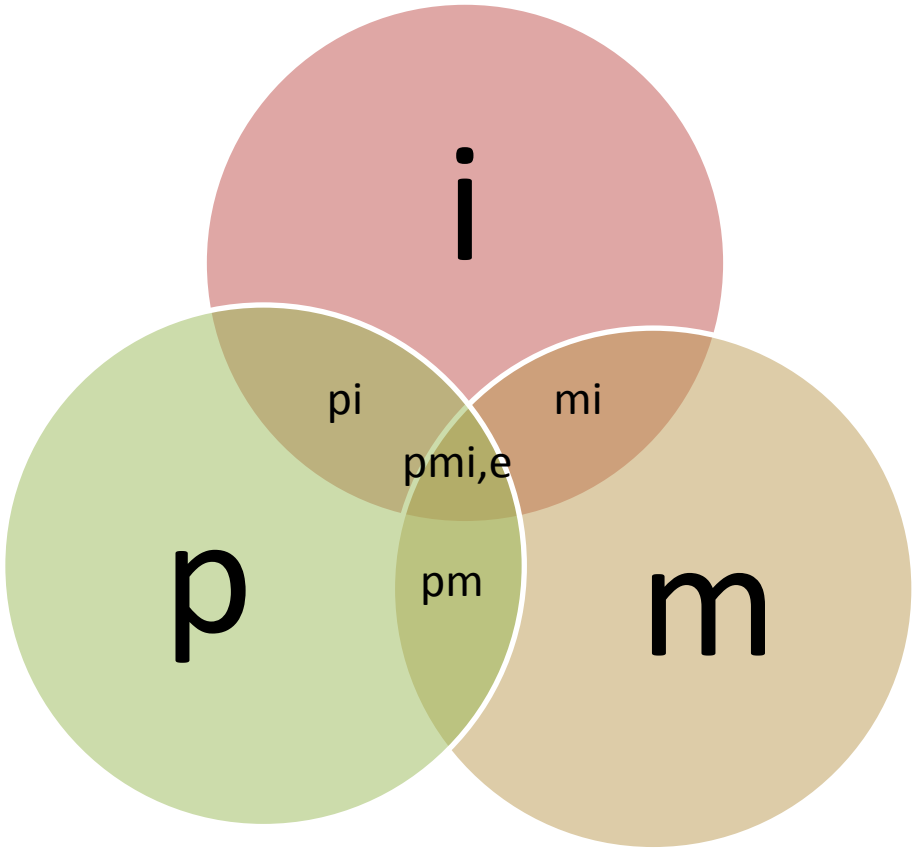
Methods for Evaluating Marking Reliability

- Variety of methods for monitoring marker consistency across marking:
  - Classical test theory
  - Generalisability theory (Brennan, 2001)
  - Many-facet Rasch modelling (e.g. Engelhard, 1996; Myford and Wolfe, 2003, 2004)
  - Hierarchical rater modelling (Patz, 1996)
  - Rater bundle model (Wilson and Hoskens, 2001)
  - Signal detection rater model (DeCarlo, Kim, and Johnson, 2011)
  - Latent trait model (Wolfe and McVay, 2012)

- Each of these methods has strengths and weaknesses

- Bejar et al. (2006) reviewed relevant literature on human marking and approaches to monitoring the quality of marking by humans

p x *m* x *i* design (script-level marking)



Sources of variability

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Variance components

$$\sigma^2(Xpmi) = \sigma^2(p) + \sigma^2(i) + \sigma^2(m) + \sigma^2(pi) + \sigma^2(pm) + \sigma^2(mi) + \sigma^2(pmi)$$

$\sigma^2(p)$ is the universal or true score variance

$\sigma^2(m)$ is the variance component for markers that estimates the between-marker variance for all admissible markers, averaging over the candidates

$\sigma^2(i)$ is the variance component for items that estimates the between-item variance for all admissible items, averaging over the candidates

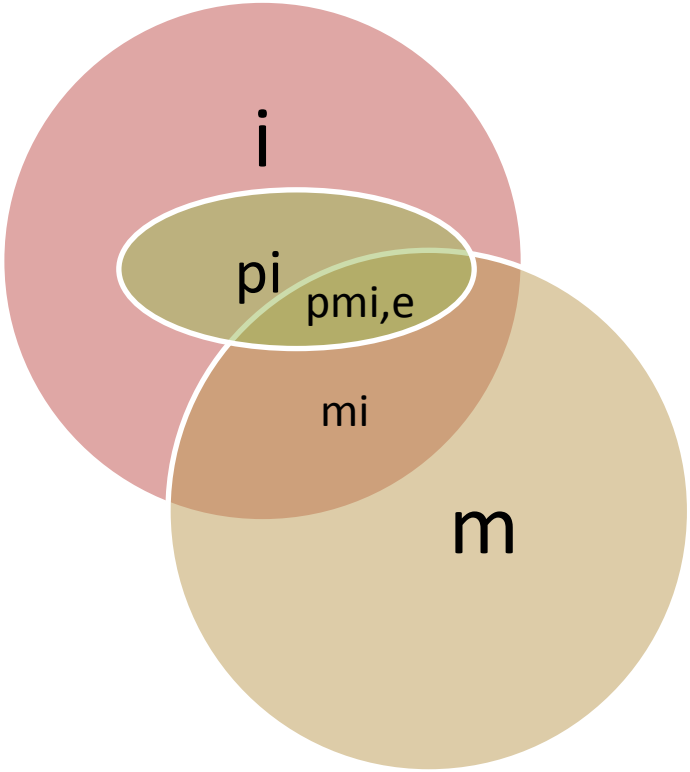
$\sigma^2(pi)$ estimates the extent to which the relative ordering of persons differs by item

$\sigma^2(pm)$ estimates the extent to which persons are scored differently by different markers

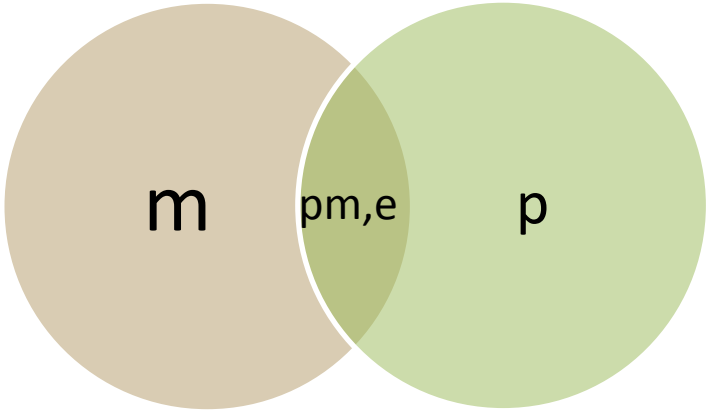
$\sigma^2(mi)$ estimates the extent to which items are scored differently by different markers

$\sigma^2(pmi)$ is the residual variance which encompasses the effect of person, item and marker interaction and other unexplained random error

p x *m* design (item-level marking)



$$\sigma^2(Xpm) = \sigma^2(p) + \sigma^2(m) + \sigma^2(pm)$$



High ‘missingness’ = analogous ANOVA

Candidate responses (seeds)	markers								
	1	2	3	4	5	6	7	8	9
1	x	x			x	x	x		x
2	x	x	x	x	x	x	x	x	x
3	x	x	x	x		x	x		
4					x			x	
5	x	x	x	x	x	x	x	x	x
6			x				x		
7	x	x	x	x	x	x	x	x	x
8			x	x	x			x	
9		x	x			x	x	x	
10		x		x	x		x		x

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, *9*, 226–252.

AQA

$$E\hat{\rho}^2(i) = \frac{\hat{\sigma}_i{}^2(p)}{\hat{\sigma}_i{}^2(p) + \left[\frac{1}{h_{mi}} - \frac{\sum_p \sum_{p'} \frac{\tilde{n}_{pp'}}{\tilde{n}_p \tilde{n}_{p'}}}{n_{pi}\left(n_{pi} - 1\right)}\right]\hat{\sigma}_i{}^2(m) + \frac{1}{h_{mi}}\hat{\sigma}_i{}^2(pm, e)}$$

$\tilde{n}_p$ is the number of markers for person *p*

$\tilde{n}_{pp'}$ is the number of markers each pair of candidates *p* and *p'* ($p \neq p'$) share in common

n_{pi} is the total number of candidates whose response to item *i* was mark-remark monitored

h_{mi} is the harmonic mean of the number of markers per candidate ($\tilde{n}_p$)

$$h_{mi} = \left[\frac{1}{n_{pi}} \sum_{p'} \frac{1}{\tilde{n}_p}\right]^{-1}$$

(see Brennan, 2001, p. 236)

$$SE\ (i) = \sqrt{\frac{1}{h_{mi}}\hat{\sigma}_i{}^2(m) + \frac{1}{h_{mi}}\hat{\sigma}_i{}^2(pm, e)}$$

Item	Marking type	Harmonic mean	Variance components						$E\rho^2$	Standard Error
			raw			as % of total variance				
			person	marker	residual	person	marker	residual		
1	Seeded	21.085	0.129	0.003	0.090	57.9	1.6	40.5	0.968	0.067
2	Seeded	19.688	0.116	0.003	0.093	54.6	1.5	43.8	0.960	0.070
3	Seeded	19.059	1.503	0.005	0.050	96.4	0.4	3.2	0.998	0.054
4	Seeded	20.697	1.161	0.011	0.046	95.3	0.9	3.8	0.998	0.052
5	Seeded	20.697	2.853	0.054	0.123	94.1	1.8	4.1	0.998	0.093
6	Seeded	23.083	0.863	0.091	0.569	56.7	6.0	37.4	0.971	0.169
7	Seeded	23.965	1.244	0.156	1.431	43.9	5.5	50.6	0.953	0.257
8	Double	2	0.757	0.350	1.170	33.3	15.4	51.4	0.504	0.872
9	Double	2	0.300	0.132	0.441	34.4	15.2	50.5	0.517	0.535
10	Double	2	2.055	0.211	1.043	62.1	6.4	31.5	0.768	0.792
11	Double	2	0.385	0.101	0.405	43.2	11.3	45.5	0.607	0.503

Item	Marking type	Original harmonic mean	$E\rho^2$			Standard Error		
			HM = original	HM = 2	HM = 1	HM = original	HM = 2	HM = 1
1	Seeded	21.085	0.968	0.734	0.580	0.067	0.217	0.307
2	Seeded	19.688	0.960	0.707	0.546	0.070	0.220	0.310
3	Seeded	19.059	0.998	0.982	0.964	0.054	0.167	0.236
4	Seeded	20.697	0.998	0.976	0.954	0.052	0.169	0.239
5	Seeded	20.697	0.998	0.970	0.942	0.093	0.298	0.421
6	Seeded	23.083	0.971	0.725	0.568	0.169	0.574	0.812
7	Seeded	23.965	0.953	0.612	0.440	0.257	0.891	1.260
8	Double	2	0.504	0.504	0.335	0.872	0.872	1.233
9	Double	2	0.517	0.517	0.346	0.535	0.535	0.757
10	Double	2	0.768	0.768	0.622	0.792	0.792	1.120
11	Double	2	0.607	0.607	0.434	0.503	0.503	0.711

Adjusting for range restriction

The marking reliability coefficient for an item *i* estimated by $E\hat{\rho}^2(i)$ is denoted by r_i.

$$R_i = \frac{kr_i}{\sqrt{k^2 r_i^2 - r_i^2 + 1}}$$

Where *k* is the ratio of standard deviations of the mark distribution of the entire item to the standard deviations of the seeded item.

Thus $k = \frac{sdu}{sdr}$ where:

sdu is the unrestricted standard deviation (i.e. standard deviations of the entire item)
sdr is the restricted standard deviation (i.e. standard deviations of the seeded items).

Thorndike, R. L. (1947). Research Problems and Techniques (Rep. No. 3 AAF Aviation Psychology Program Research Reports). *Washington, DC: US Government Printing Office*.

The future (is now)

- *p* x *m* x *i* design for unit-level marked components/ aggregation to unit-level for item-level components
- Aggregation to specification level
- Sawtooth plots (of grade confidence)

- Targeted review of the worst items and specifications
- How to resolve the seed representativeness issue?
  - Particularly as data is from a primarily operational system…