Ydy itemau wedi eu cyfieithu yn perfformio yn debyg?

Do translated items perform the same way?

Profiad asesu mewn gwlad ddwyieithog

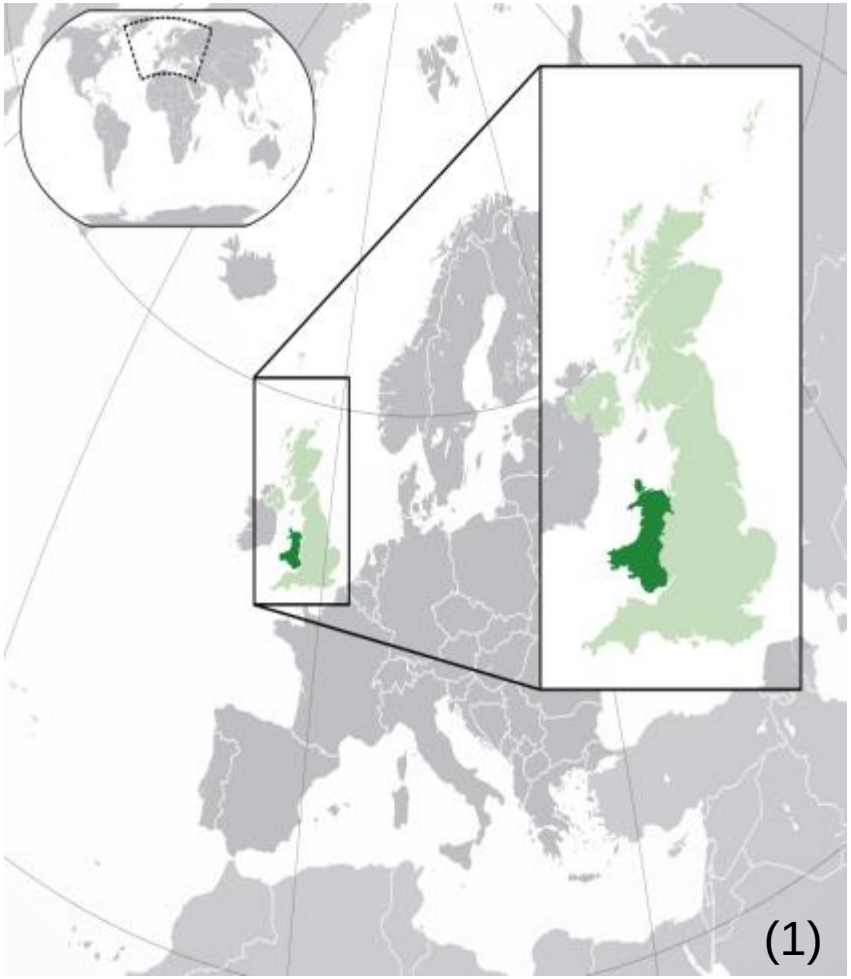
The experience of assessment in a bilingual country.

Mark Hogan, Statistician
November 2017

Version Control

Version	Author	Date	Notes
1	Mark Hogan	01 Sep 17	First Draft.
2	Mark Hogan	06 Sep 17	Included Su17 WJEC GCE AS Economics example.
3	Mark Hogan	13 Sep 17	Amended following Translation feedback.
4	Mark Hogan	04 Oct 17	Included Evans relationship FF plot and MD results for Su17 GCE AS Economics example.
5	Mark Hogan	12 Oct 17	Corporate formatting.
6	Mark Hogan	16 Oct 17	Updated following comments from Translation.
7	Mark Hogan	18 Oct 17	Updated following comments from RH, GP and EC.
8	Mark Hogan	19 Oct 17	Updated following AE comments.
9	Mark Hogan	20 Oct 17	Minor improvements.
10	Mark Hogan	27 Oct 17	Minor edits.
11	Mark Hogan	30 Oct 17	Edits following GP feedback.

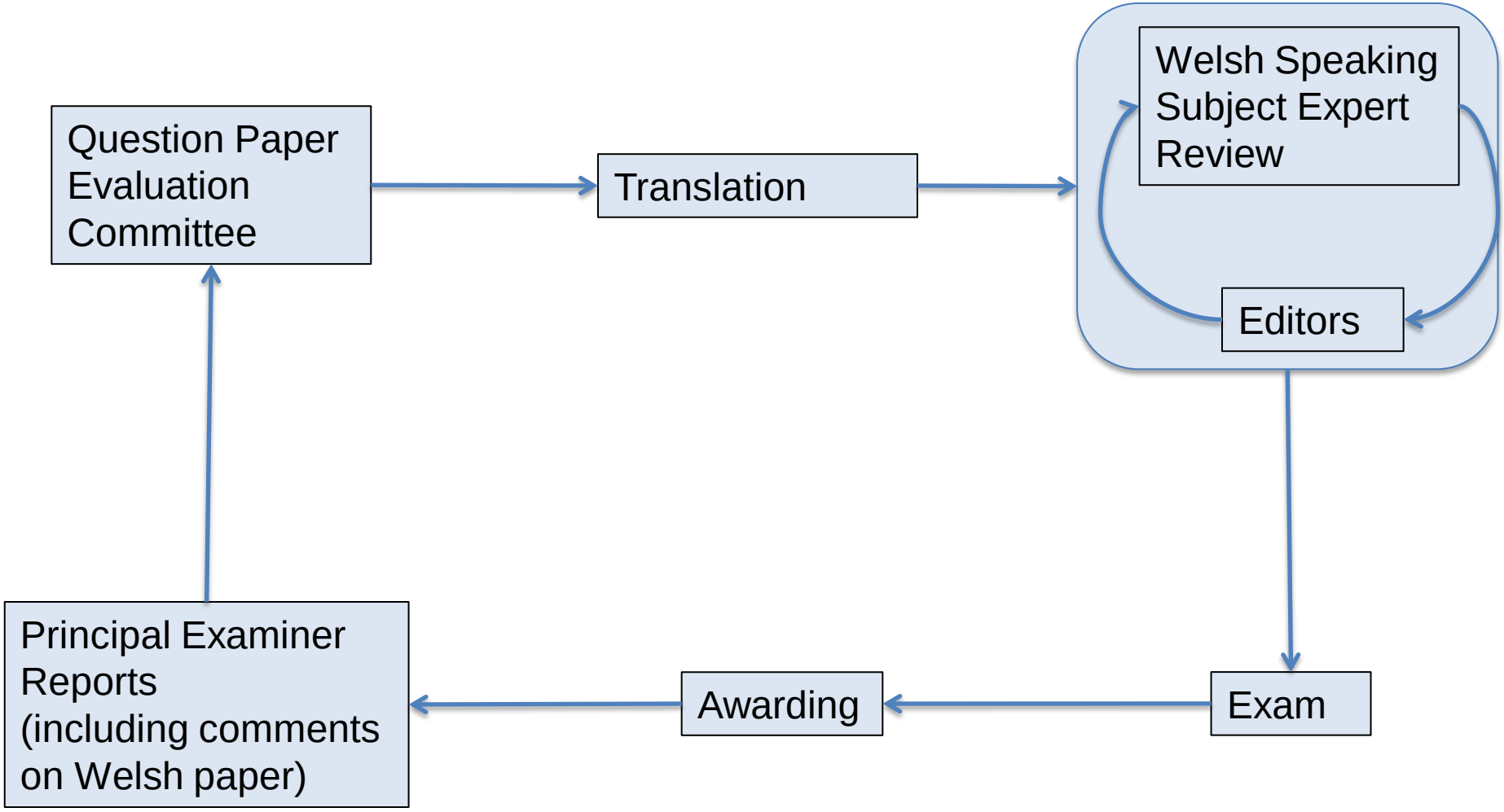
Cefndir :: Background



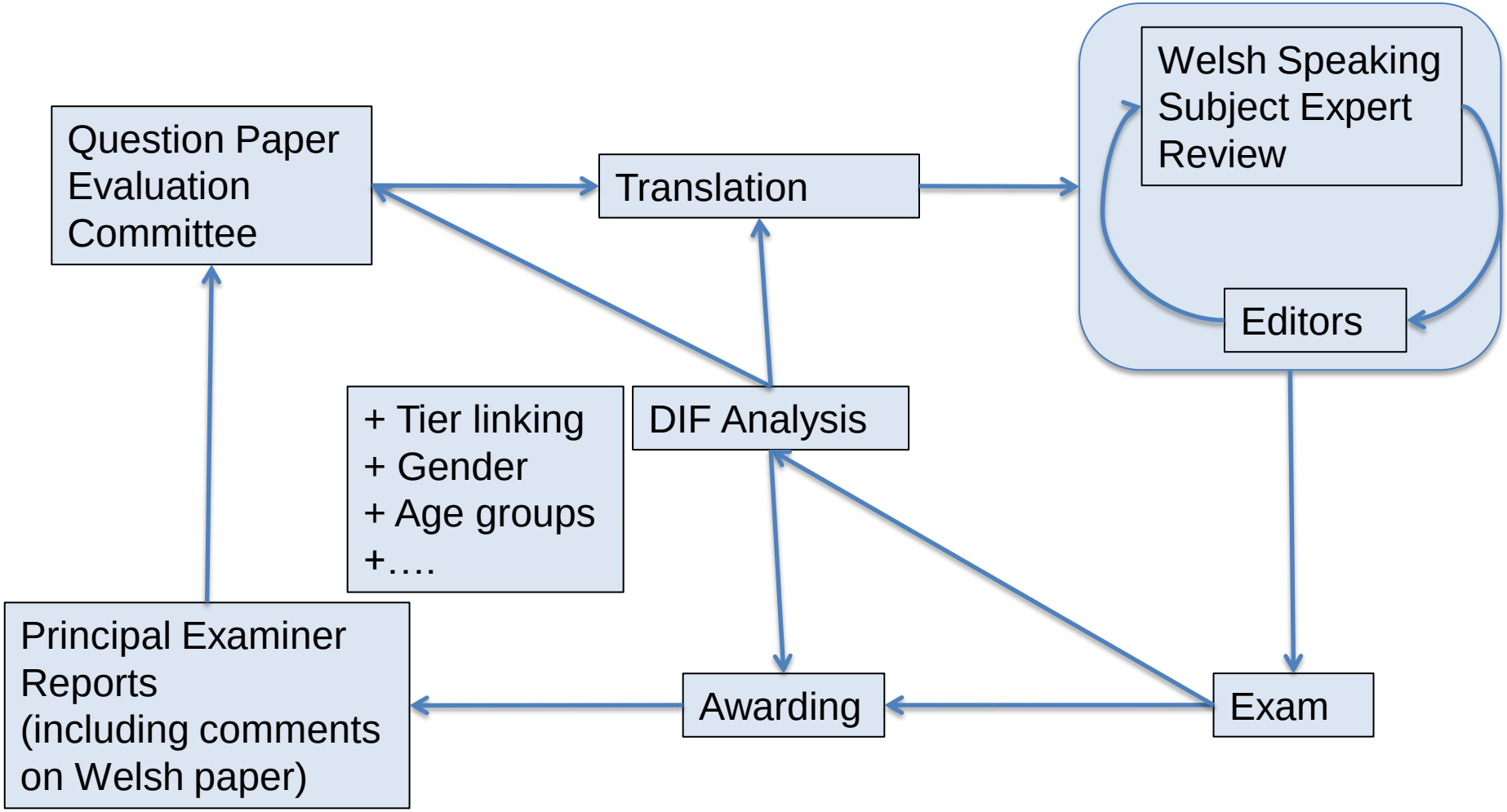
- Poblogaeth 3.1 miliwn (2)
- 19% yn siarad Cymraeg (2)
- ymgeiswyr yn cael dewis sefyll trwy gyfrwng y Gymraeg neu'r Saesneg
- Mae tua 14% o ymgeiswyr sefyll trwy cyfrwng Cymraeg.
- CBAC yw prif fwrdd arholi Cymru
- tua 30,000 ymgewiswyr TGAU
- tegwch – dim DIF rhwng cyfrwng iaith

- 3.1 million population (2).
- 19 % can speak Welsh (2).
- Candidates may chose to sit exam in medium of either English or Welsh.
- Approximately 14 % of candidates sit Welsh medium paper.
- WJEC is Wales' main Examination Board.
- Circa 30,000 GCSE candidates.
- Fairness – no DIF between language medium.

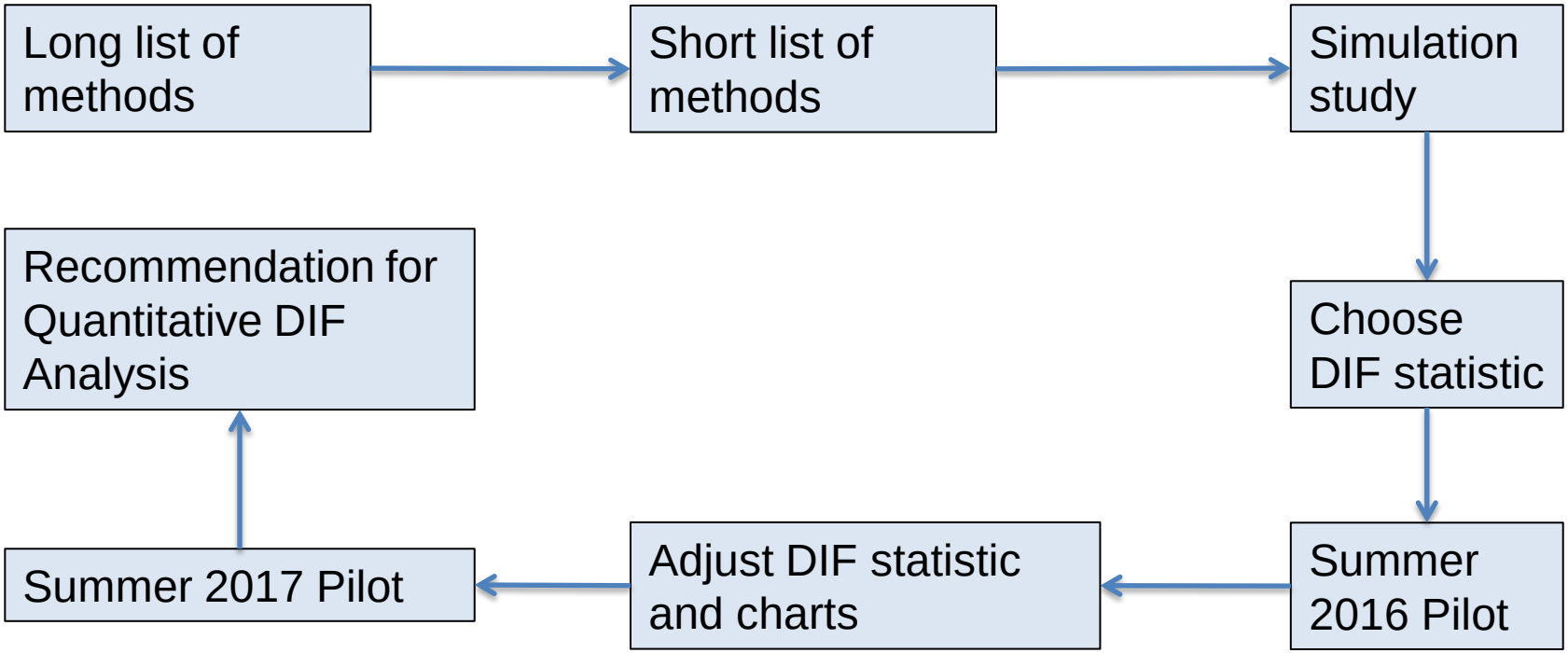
Exam Paper Translation - Current



Exam Paper Translation - Proposed



Research Method



DIF Method

Short List (building on Wiberg, 2007 (3))				
Method	**Polytomous Items**	**Measure DIF**	**Test DIF**	**Uniform**
MD	✓	✓	✓	✓
SMD	✓	✓	✓	✓
Mantel-Haenszel	✓	✓	✓	✓
Logistic Regression	✓	✓	✓	✓
IRT – Exact Unsigned Area	✓	✓	✓	✓
Simple Unsigned Area	✓	✓	X	✓
Simple Signed Area	✓	✓	X	X
IRT – Likelihood Ratio	✓	X	✓	✓
Log Linear Modelling	✓	X	✓	✓
Graphical Item Response Function	✓	X	X	✓
Chi-square Methods	X	X	✓	✓
IRT – Lord's Chi-Square	X	X	✓	✓

DIF Method

Simulation Study	
Method	**Results**
MD	• Lowest false positive rate • Good detection above DIF of 0.5 marks
SMD	• High false positive rate • Good detection above DIF of 0.5 marks
Mantel-Haenszel	
Logistic Regression	
IRT – Exact Unsigned Area	• Not tested due to requirement of large sample sizes and software limitations.

Pilot – Summer 2016 GCSE - Outputs

- Penfield and Camili strategy (4)
  1. Establish stratifying variable
  2. Determine reliability of the stratifying variable
  3. Measure between-group differences in target trait distribution
  4. Compute DIF statistics
  5. Conduct content analysis of items flagged as having DIF
  6. Remove large DIF items and recompute DIF statistics
- Boxplot and entry sizes
- Item v Rest Score
- Mean interitem correlation, item rest correlation, item test correlation, reliability
- Ability distributions
- Facility Factor scatterplot (similar to Angoff's delta plot (5))
- Item rest mean with CI
- Standardized Mean Difference with p-value

Pilot – Summer 2016 GCSE - Feedback

Subject Officers

- Welsh paper, bilingual or English responses.
- Flexible English paper provision.
- Target age group, language maturation (RS birth control, euthanasia)

Translators

- Welsh speaking subject expert checks for majority
- Difficult Welsh word -> Include English in brackets
- Include number of centres and centre details
- Extensive work over the last 10 years
  - National effort to standardise terminology (Y Termiadur Addysg)
  - Improved Welsh medium Teacher training
- FF plot most useful
- High false positive rate
- Fine recording of item and sub-item marks

Pilot – Summer 2017 GCSE

- Switched to Mean Difference (MD)
- Used facility factor plot with confidence intervals (derived from log-log relationship, allows for non-uniform DIF)

Example

Question

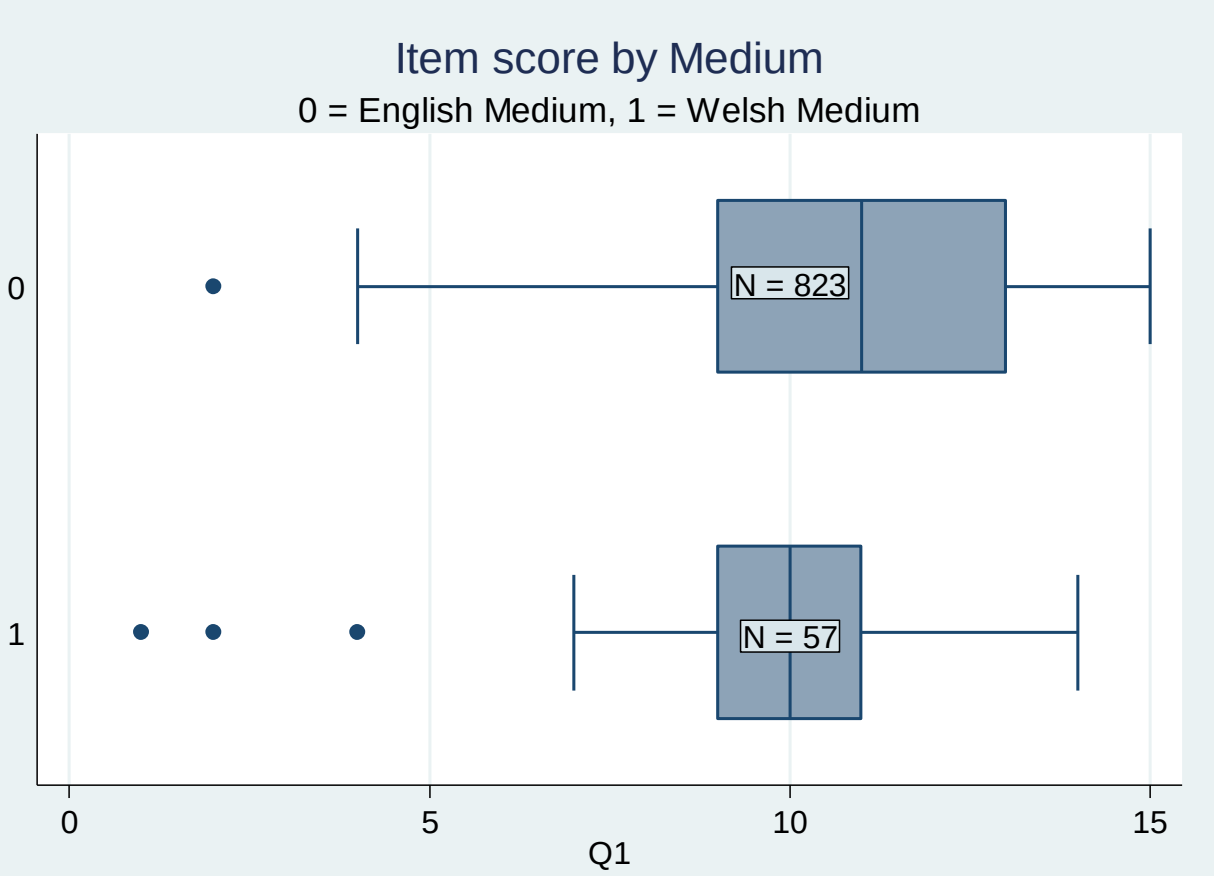
- Response Option 1
- Response Option 2
- Response Option 3
- Response Option 4
- Response Option 5

Cwestiwn

- Opsiwn Ymateb 1
- Opsiwn Ymateb 2
- Opsiwn Ymateb 3
- Opsiwn Ymateb 4

Opsiwn Ymateb 5

Example

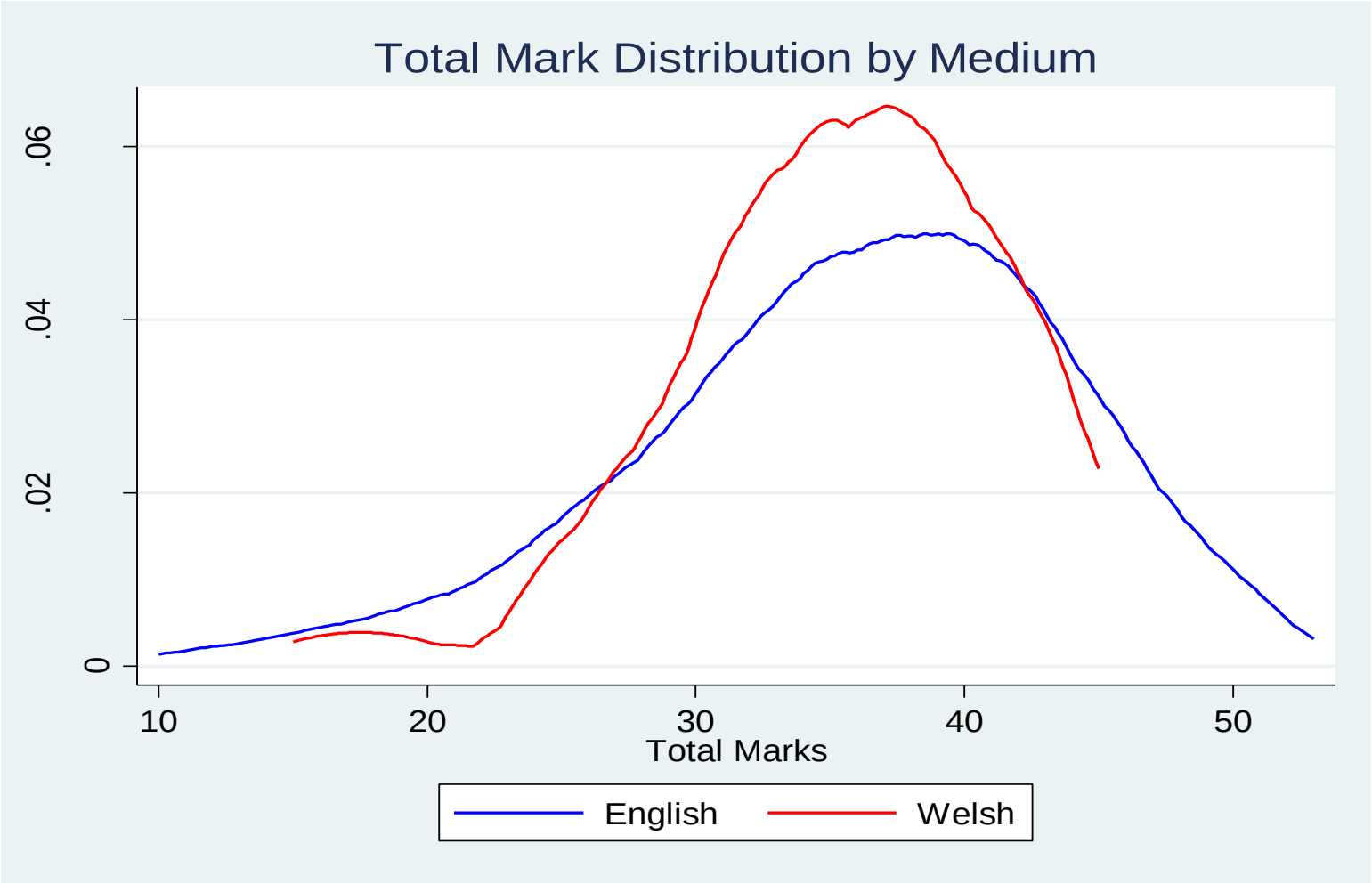


Welsh Medium Entries		
School	**English**	**Welsh**
A	0	6
B	0	9
C	0	20
D	0	22

Example

Item Reliability Statistics					
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Mean Inter-Item Correlation	0.36	0.41	0.35	0.32	0.36
Item Rest Correlation	0.50	0.38	0.52	0.61	0.51
Item Test Correlation	0.70	0.61	0.71	0.77	0.70
Alpha	0.69	0.74	0.69	0.65	0.69

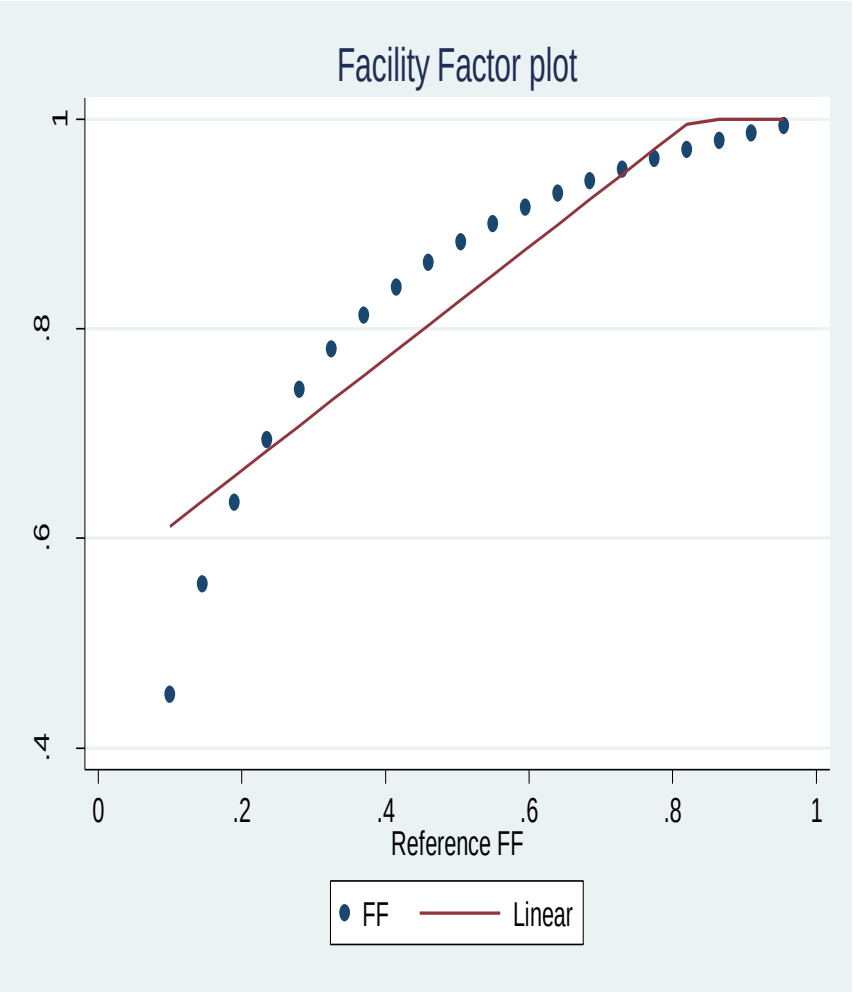
Example


Total Mark Distribution by Medium
.06
.04
.02
0
10
20
30
40
50
Total Marks
English
Welsh

Example

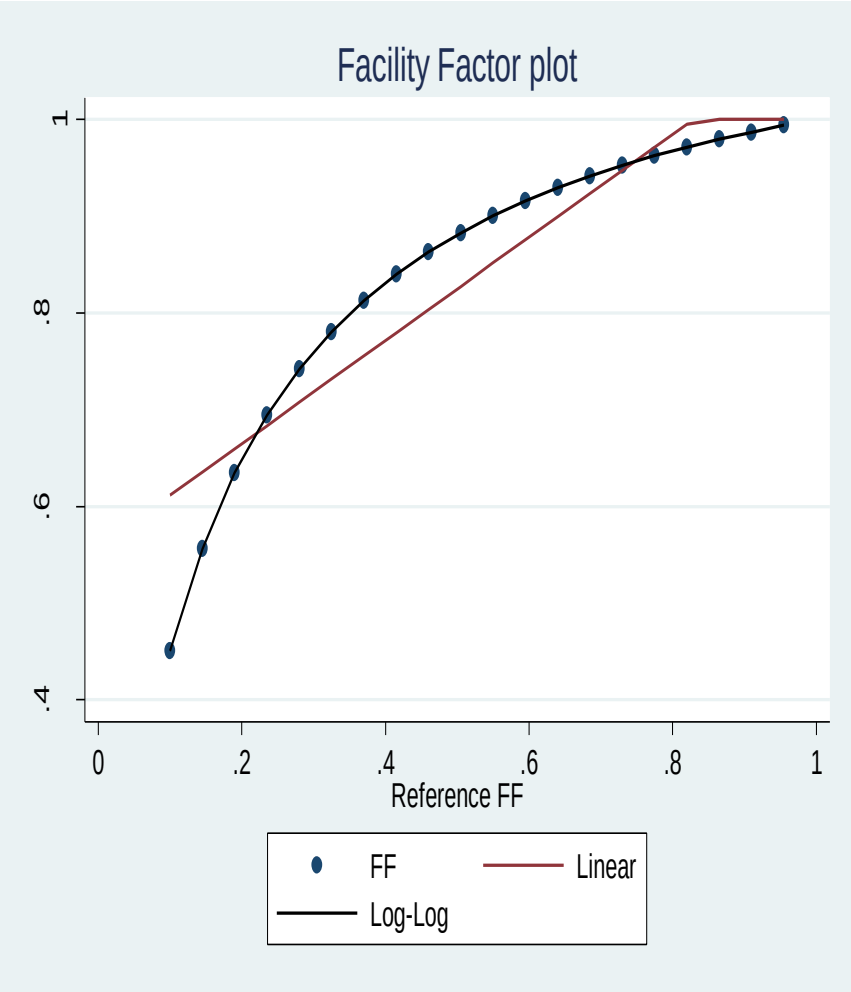

Item Score v Rest Score
Q1
0
5
10
15
0
10
20
30
40
Rest Total
Line best fit Eng.
Line best fit Wel.

An aside – FF relationship



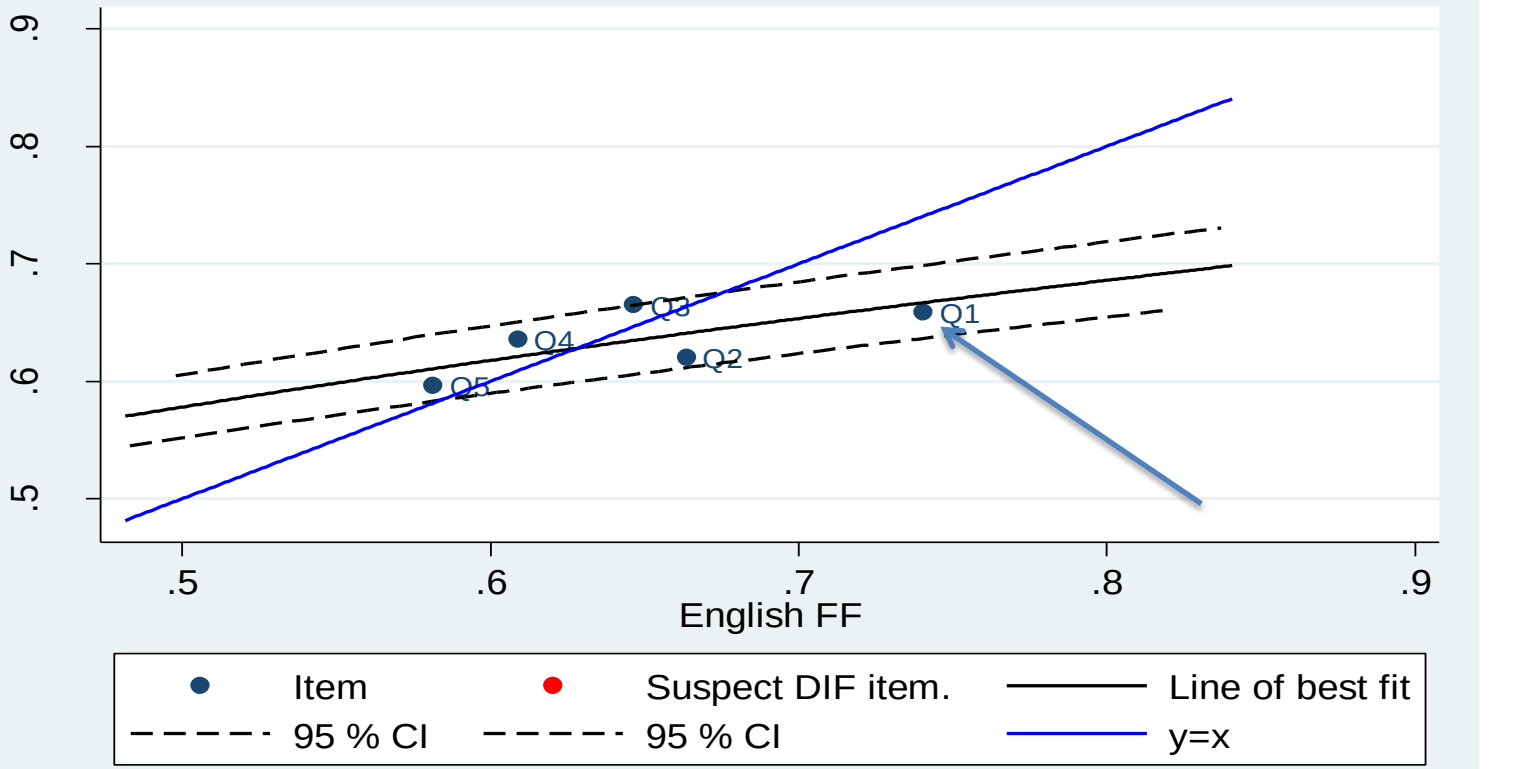
- FF plots often assume linear relationship.
- Analysis suggested non-linear.
- Mathematical derivation of expected relationship assuming:
  - Dichotomous items;
  - One-parameter Rasch model.

An aside – FF relationship



- FF plots often assume linear relationship.
- Analysis suggested non-linear.
- Mathematical derivation of expected relationship assuming:
  - Dichotomous items;
  - One-parameter Rasch model.
- Result
  - Reduces false positives at extremes and middle.
  - Second order approximation not successful.
  - Log-log regression worked well.
  - Use orthogonal regression, residuals and CIs.
  - Allows for non-uniform DIF.

Example


Facility factor plot by language (R2 = .291).
Above line of best fit = easier for Welsh, below line of best fit = harder for Welsh
.5
.6
.7
.8
.9
Q1
Q2
Q3
Q4
Q5
English FF
Item
Suspect DIF item.
Line of best fit
95 % CI
95 % CI
y=x
Orthogonal log-log regression. N Eng = 823, N Wel = 57. Wel-Eng~N(-.01,.05)

Example

DIF Analysis Summary	
Analysis	**Results**
Boxplot and Entries	Welsh candidates score slightly lower. N. B. Small Welsh entry.
Reliability	Suspect DIF item has similar correlations. Scale is reliable. DIF analysis can be performed.
Ability Distribution	Similar ability distribution for candidates. DIF analysis can be performed.
Facility Factor Plot	No cause for concern. Suspect DIF item is within 95 % CI and very close to line of best fit.
Item-Rest Plot	Welsh candidates seem to achieve fewer marks (uniform DIF), however small Welsh entry provides poor statistical power.
MD	Welsh candidates achieve 1.4 fewer marks, strongly statistically significant (p-value < 0.01). Statistical power is 0.83.
Quantitative Analysis Result	Small Welsh entry weakens statistical evidence strength. Item does not have unusual DIF compared to other items in the paper and there is inconsistent statistical evidence (FF plot insignificant, MD significant). Therefore move to qualitative review but note weak and inconsistent statistical evidence.
Qualitative Review Outcome	Reviewed by Awarding Committee and Subject Officer and conclusion was minimal or no effect and so no adjustment was justified.

DIF Analysis – Next Step Options

Qualitative review by Translators, Editors and Welsh Speaking Subject Experts.

Options following Qualitative Review	
Option	**Notes**
Do nothing	• Qualitative review finds no issue with suspected DIF item and concludes the DIF analysis result was a false positive.
Adjust marks of affected candidates	• Qualitative review finds an issue with the item. • Requires careful consideration of choosing uniform or non-uniform DIF analysis.
Drop item for awarding purposes.	• Qualitative review finds an issue with the item. • Extreme last resort. How does this affect the underlying construct, assessment objectives and candidate outcomes?

Crynodeb :: Summary

- Argymhelliad.
  - Gwahaniaeth Cymedrig gydag allbynnau graffigol.
  - Plot FF yn defnyddio perthynas bras log-log ac atchweliad orthogonal.
  - Cofnodi diwedd marciau eitem ac is-eitem.
- Dadansoddiad DIF yn gam cyntaf, adolygiad ansoddol yn gam nesaf.

- Recommendation
  - Mean Difference with graphical outputs.
  - FF plot uses log-log relationship approximation and orthogonal regression.
  - Finer recording of item and sub-item marks.
- DIF analysis only first step, next step is qualitative review.

Gwaith i’r dyfodol :: Future Work

- Adborth
  - QPEC a Chyfieithu (positifau ffug)
  - ymholiadau (negatifau ffug)
  - Meini prawf arwyddocaol / adolygiad ansoddol
- Ystadegyn DIF Anffurfiol?
- Sut i ddelio ag eitemau dewisol?
- Goblygiadau ar gyfer DIF ar gyfer rhyw, oedran, ac ati.
- Diddordeb rheoleiddiol mewn DIF yn ôl gallu.
- Prawf Gwahaniaethol / Swyddogaeth Cymhwyster.

- Feedback
  - QPEC and Translation (false positives)
  - enquires (false negatives)
  - Significance/Qualitative review criteria
- Non-Uniform DIF statistic?
- How to deal with optional items?
- Implications for DIF for gender, age, etc.
- Regulatory interest in DIF by ability.
- Differential Test/Qualification Functioning.

References

1. Wikipedia,2017. *Wales in the UK and* Europe. [online] Available at: https://commons.wikimedia.org/w/index.php?curid=18497747 [Accessed 20 October 2017].
2. Office for National Statistics, 2012. *2011 Census: Key Statistics for Wales, March 2011*. [online] Available at <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/2011censuskeystatisticsforwales/2012-12-11#proficiency-in-welsh> [Accessed 20 October 2017].
3. Wiberg, M., 2007. *Measuring and Detecting Differential Item Functioning in Criterion-Referenced Licensing Test, A Theoretic Comparison of Methods*. [pdf] Umea University. Available at: <http://www.edusci.umu.se/digitalAssets/59/59534_em-no-60.pdf> [Accessed 12 October 2017].
4. Penfield, R. D., Camilli, G., 2007, Differential Item Functioning and Item Bias. In: Rao et al, 2007, *Handbook of* Statistics, vol 20,pp. 125-167.
5. Magis, D., Facon, B., 2012, *Angoff's delta method revisited: Improving DIF detection under small samples*, British Journal of Mathematical and Statistical Psychology, vol 65, issue 2,