Detecting Differential Rater Functioning

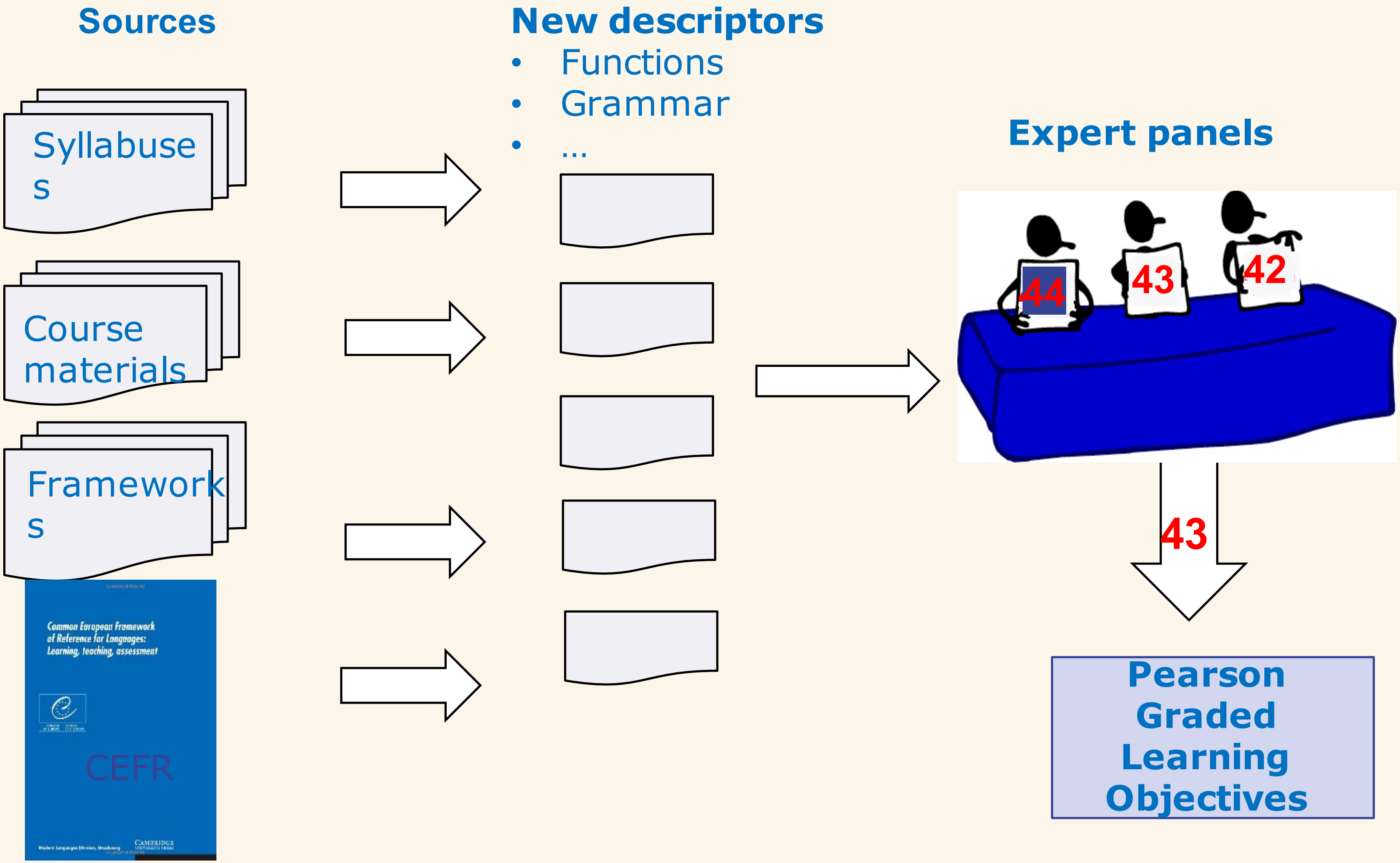
John de Jong

Daeryong Seo

Durham,18 March 2016

Adding Descriptors: Calibrating learning objectives

Expert judges assign GSE values to learning objectives



Sources

Syllabuses

Course materials

Frameworks

Common European Framework of Reference for Languages: Learning, teaching, assessment
CEFR
Cambridge

New descriptors

- Functions
- Grammar
- …

Expert panels

Pearson Graded Learning Objectives

Example: developing new descriptors

- Write ±100 new descriptors (Can-do statements)
- Rate descriptors
  - 89 experienced course-ware developers from Pearson based in 10 countries across the world assigned GSE scale values (10 – 90)
  - 316 teachers from ± 50 countries with a detailed knowledge of the CEFR and a minimum of two years teaching experience classified the descriptors at one of the CEF levels (<A1 – C2)
  - The average classifications from the teachers were then projected onto the GSE
- Average ratings from the two independent groups correlated 0.981

The Global Scale of English and the CEF

The Global Scale of English **is linked** to the CEF at various anchor points through

- Alignment of tests
- Standard setting using samples of written and spoken production
- Inclusion of CEF descriptors in syllabus calibration process

The Global Scale of English **complements** the CEF by providing

- Coverage of skills and levels where CEF descriptors are sparse
- Detailed description in relation to learning objectives for English

GSE Events and Activities

TESOL 2015 (Toronto, Canada)
BAAL SIG in Language Learning & Teaching (Edinburgh) [GSE vocabulary]
Polish IATEFL (Krakow)
KOTESOL (Seoul, Korea) [launch of GSE Academic Learning Objectives]
British Embassy (Tokyo, Japan)
British Council conference (Seoul, Korea)
MEXTESOL (Cancun, Mexico)
BESIG (Sitges, Spain) [launch of GSE Professional Learning Objectives]
JALT (Shizuoka, Japan)
LTF (Oxford, UK)
Language Assessment Conference (Guangzhou, China)
Warsaw University of Technology (Warsaw, Poland)

Content & Assessment

15 courses now mapped and badged with GSE ranges

External mapping: Malaysia and Japan

Polish Matura alignment: student data collection complete, standard setting complete. Results being analysed

Cambridge First: pilot study underway to test feasibility

6 levels of Progress published; one level as a BETA

Enhancements to Advisory notices for Progress implemented

Initial live data for Progress shows tests working, for the most part, as expected

PTE Academic has grown significantly; increase in enquiries on results especially around speaking - work ongoing to explore this

Further standard setting studies probable including against Canadian Benchmarks

IRT 1 & re-rating

Outcomes IRT1:

GSE values for Adult LOs - satisfactory. Published externally.

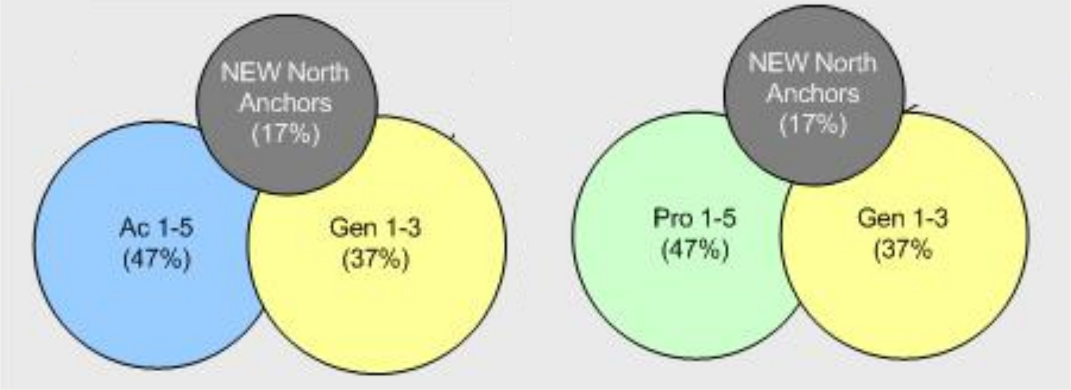
GSE values for Young Learner LOs - satisfactory but too few to publish

GSE values for the Academic & Professional LOs - lower than expected,artificial cap at ~75. Planned public release delayed for investigation.

Re-rating

140 existing B1-C2 descriptors sent for re-rating with 14 additional C1/C2 North 2000 anchors

~ 700 new LOs rated and classically analysed



IRT 2

IRT2 Data

12 batches from IRT1

7 new batches

2 ‘special’ batches of re-rated data

Total LOs = 2001

Total raters = 7599

IRT2 & calibration method

Free-calibration, 1-parameter model using WINSTEPS

4 runs - data evaluated after each run. Total of 699 raters and 158 LOs removed for misfit

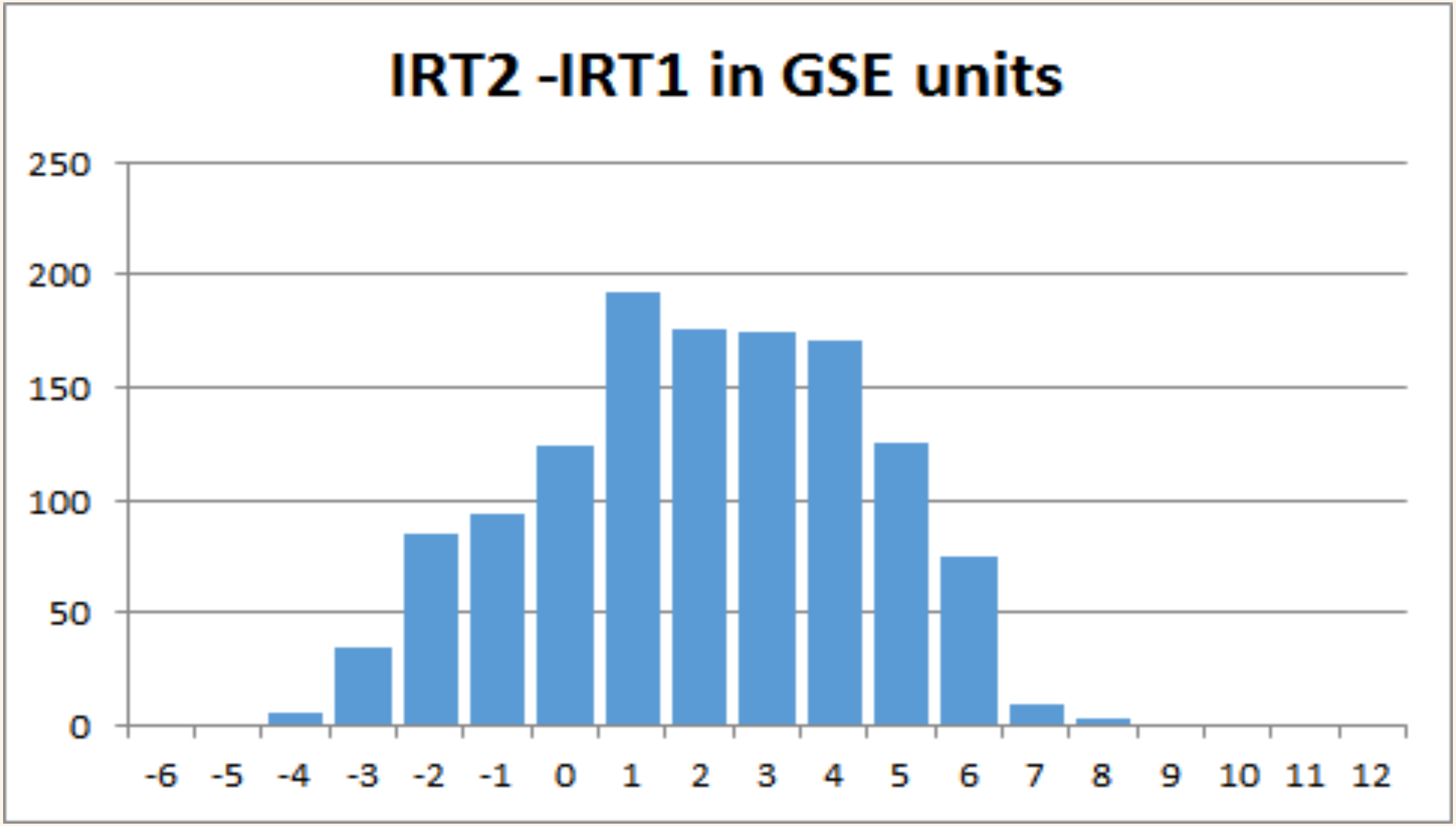
e.g. N<80, INMSQ/OUTMSQ >2.56, Count<25 and pointbiserial <0.10,

IRT estimates were transformed onto the North 2000 scale using a linear regression.

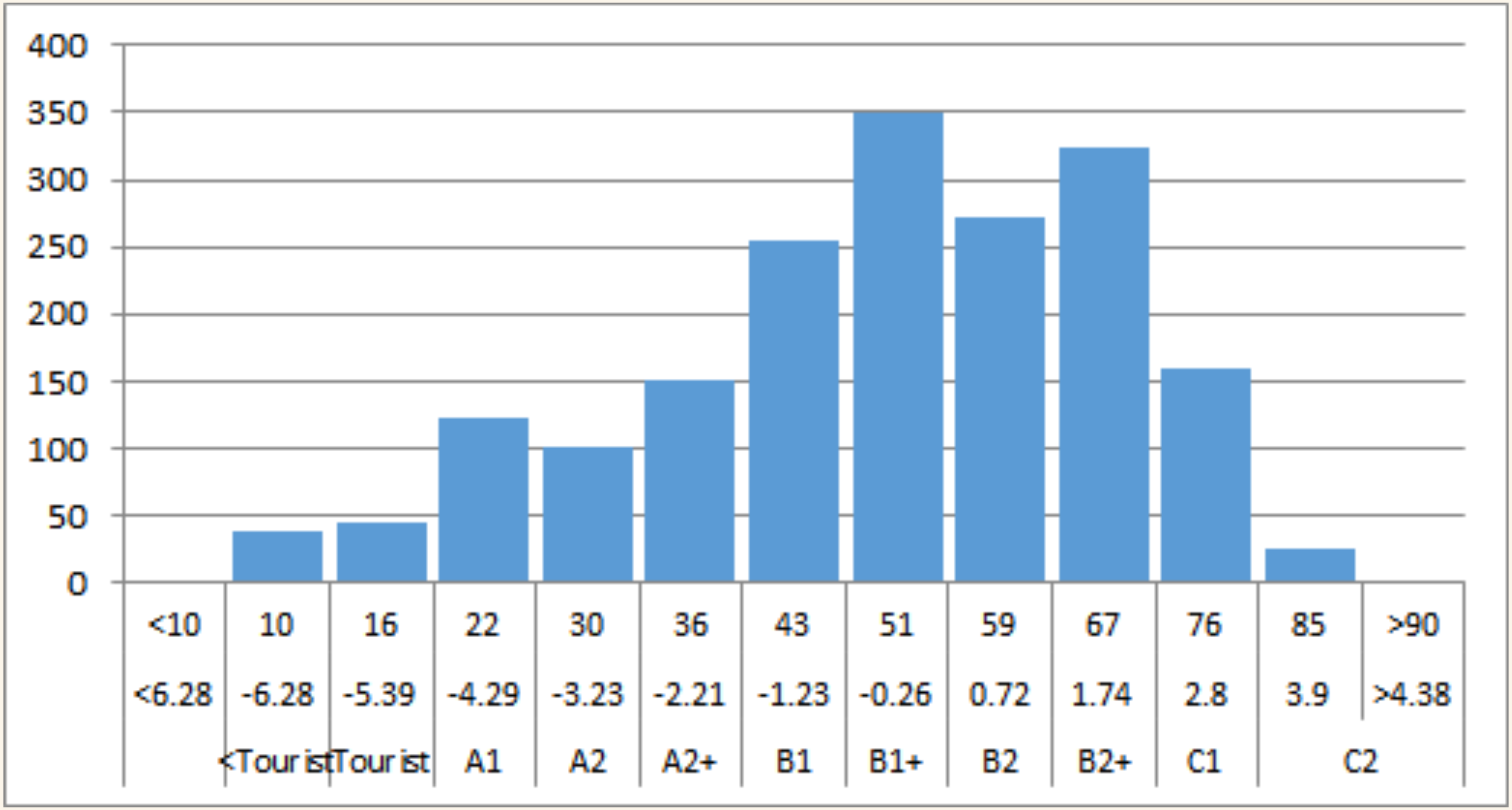
GSE values were then created for each descriptor by converting the North 2000 scale to GSE using a transformation function

Some shift to IRT1 outcomes

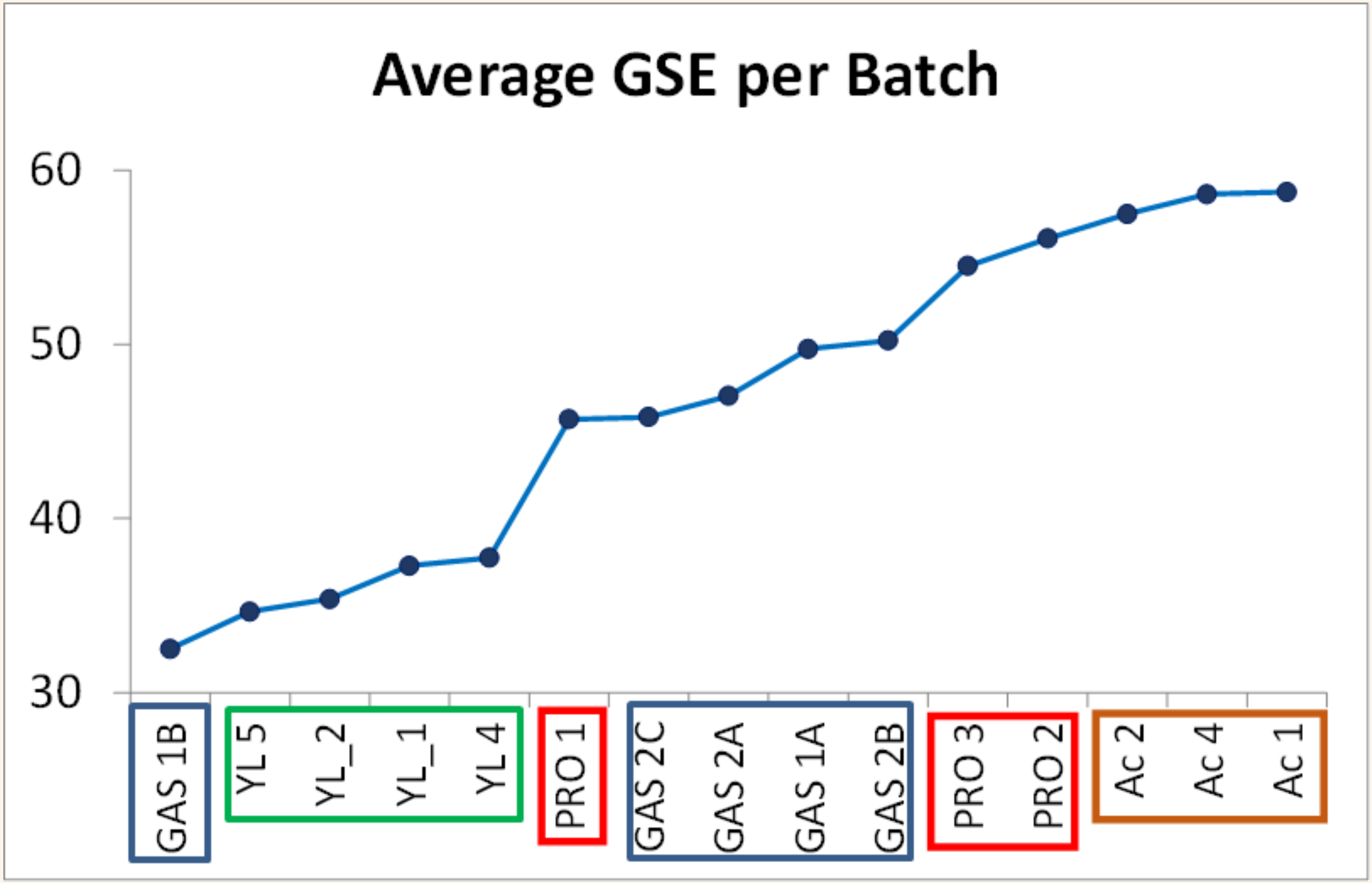
IRT2 Outcomes: changes



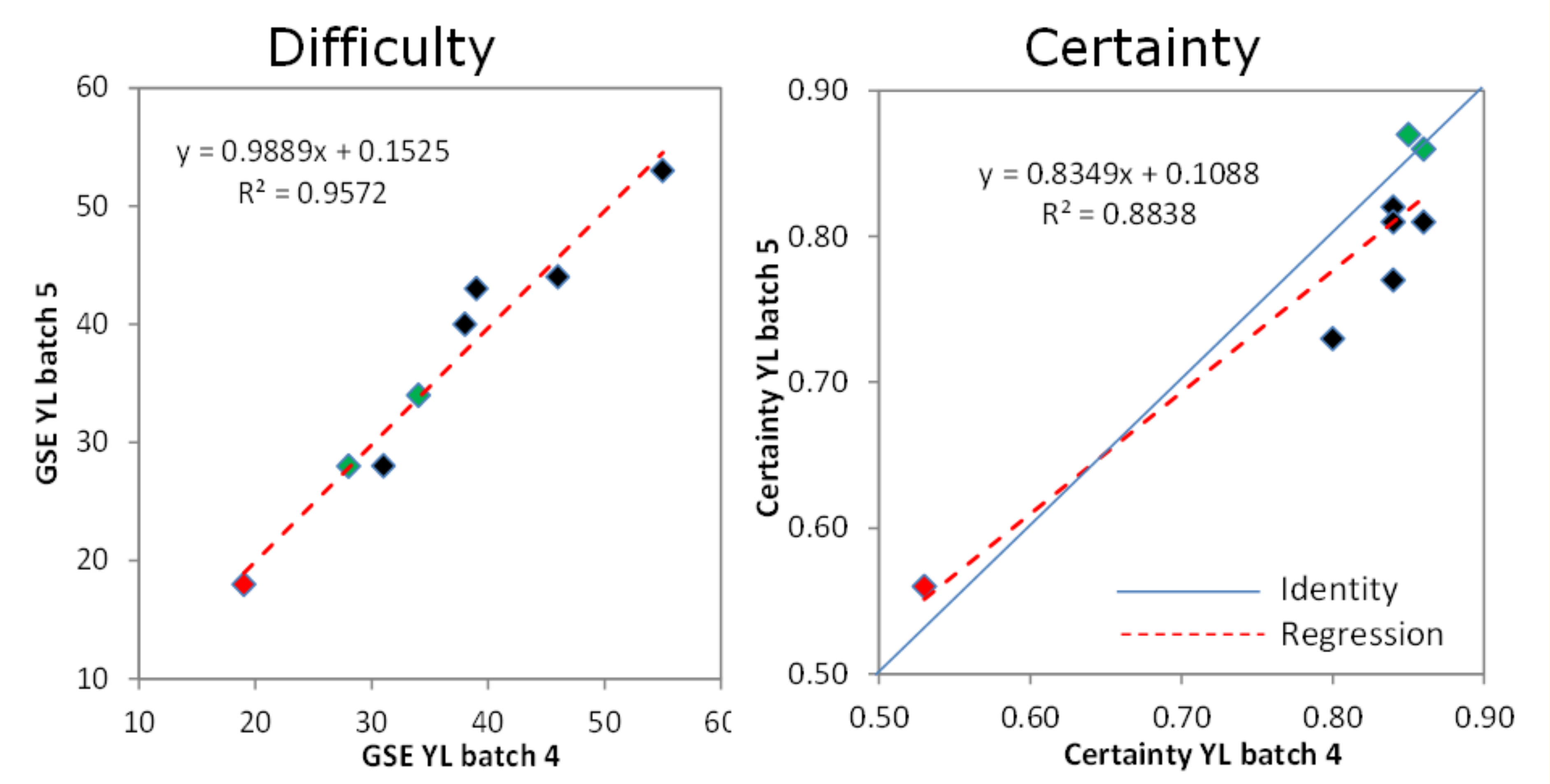
Distribution of Learning Objectives



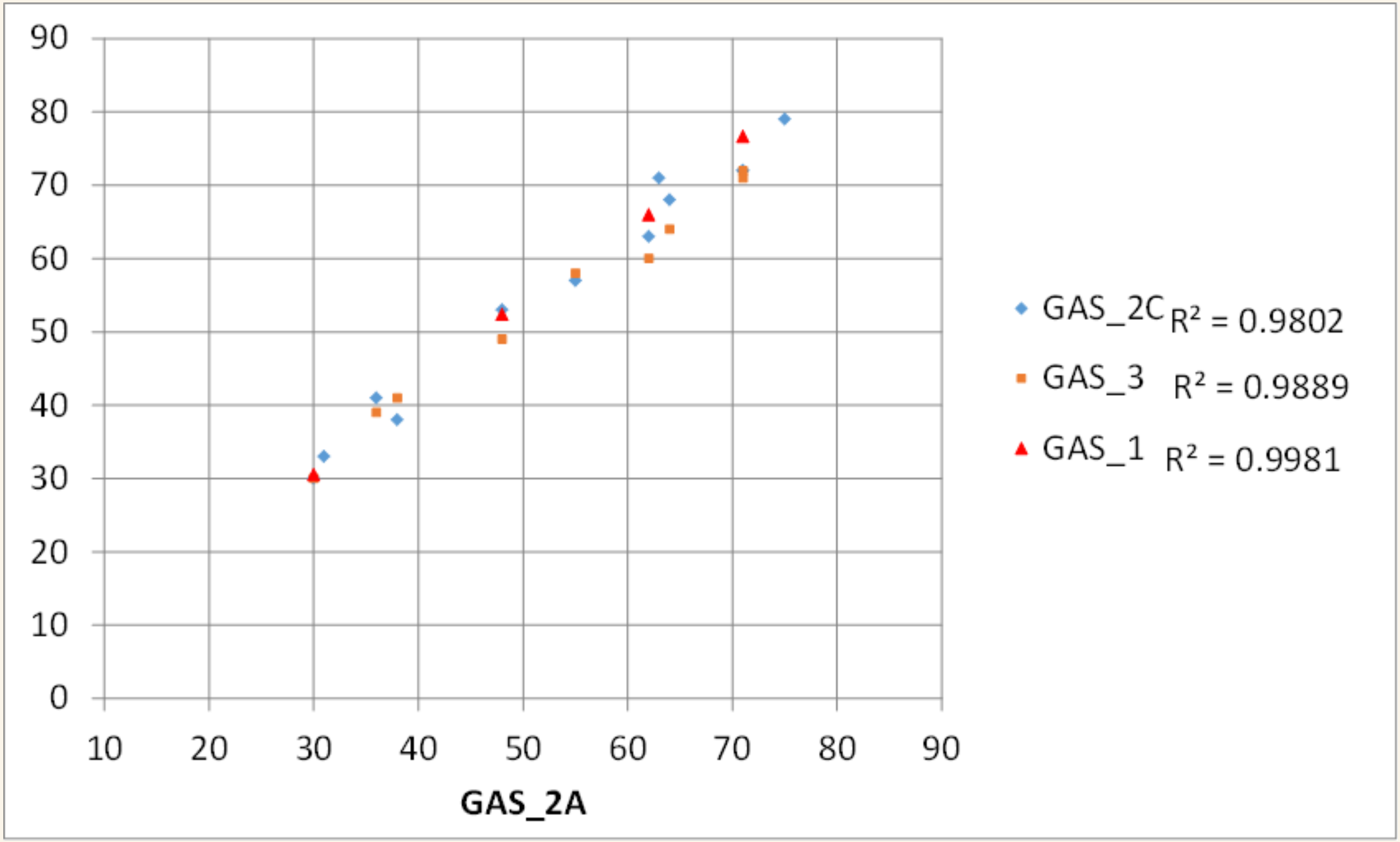
Average GSE per Batch



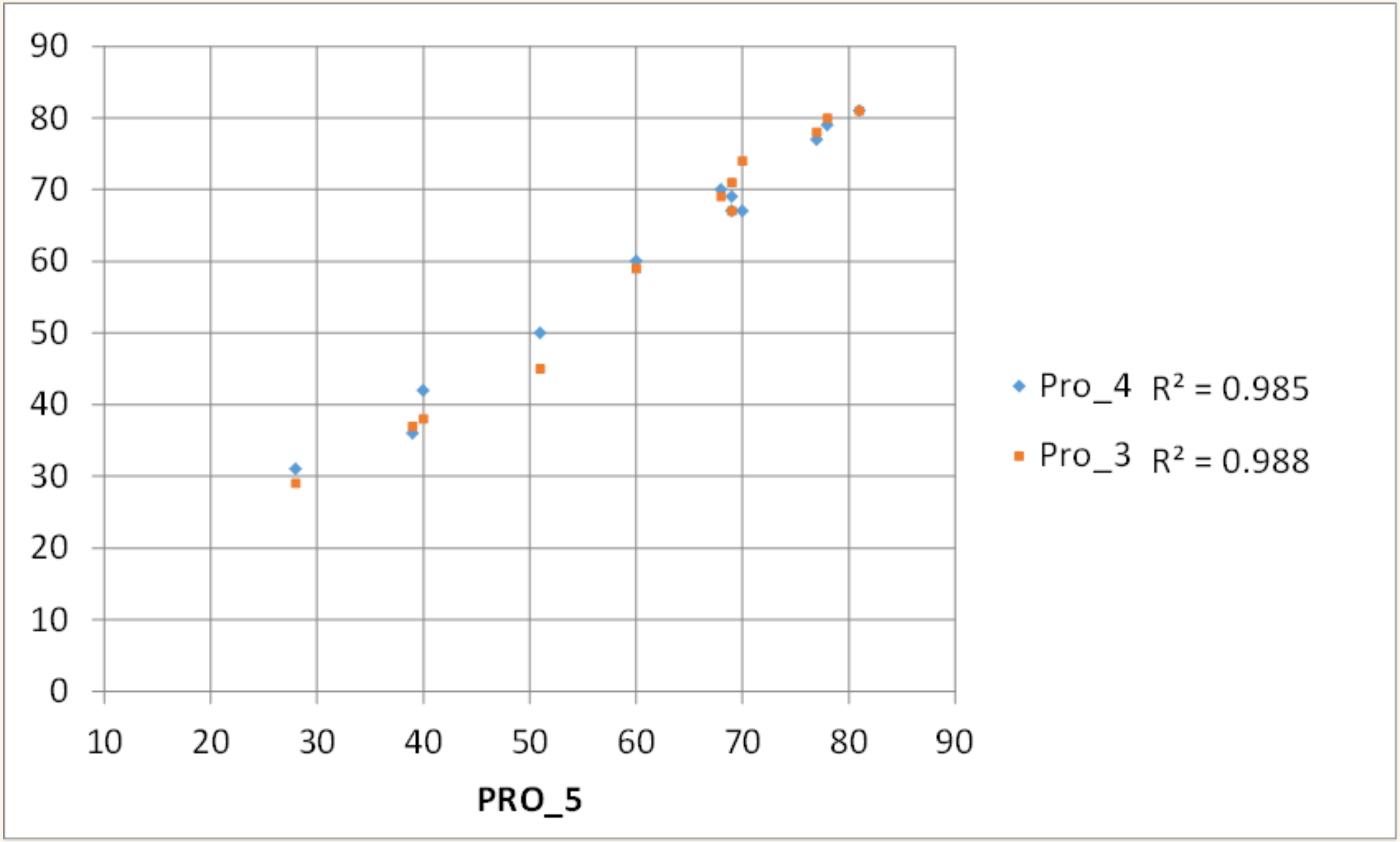
North Descriptors in YL Batches 4 & 5



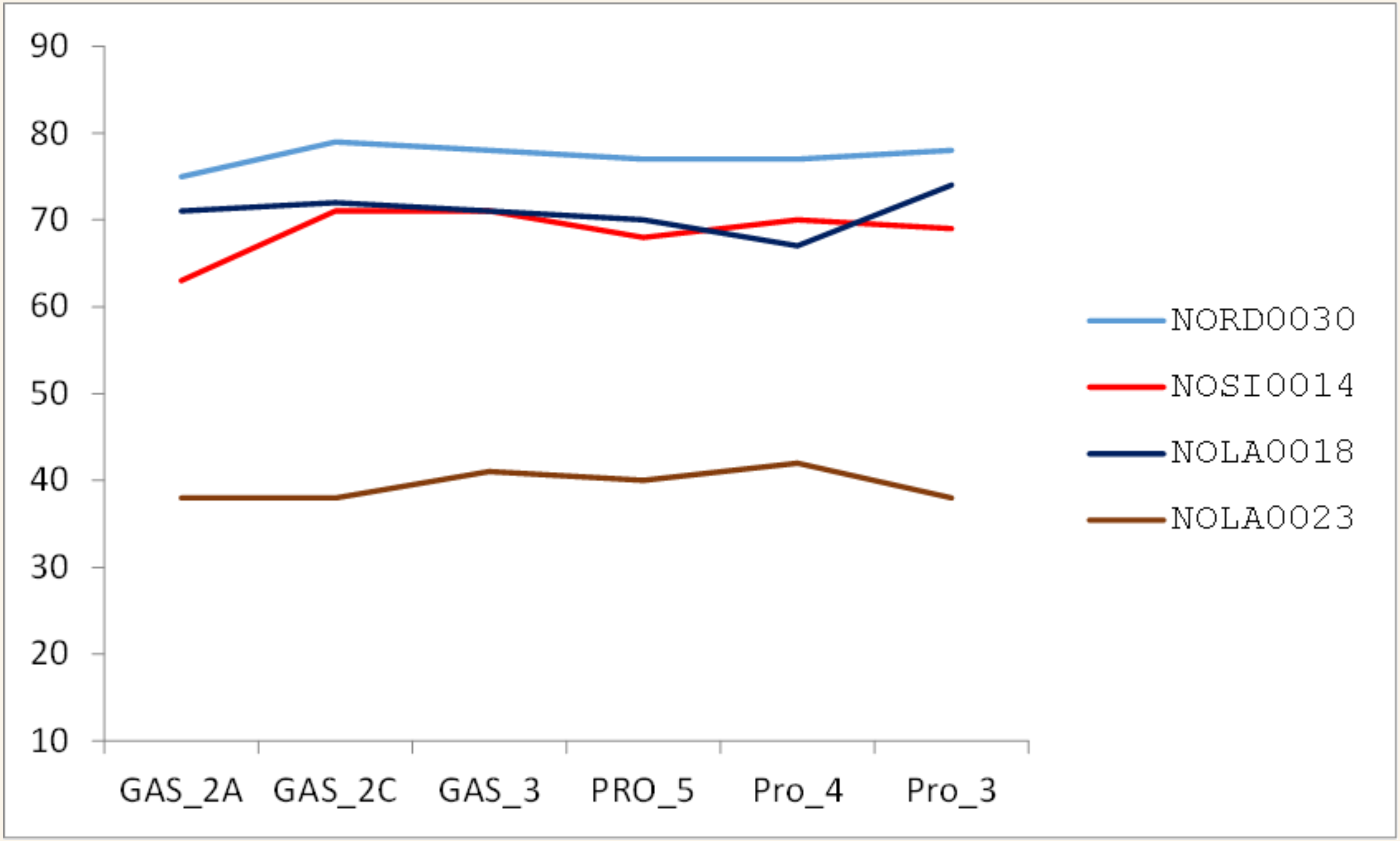
Stability of North Anchors in 4 GAS Batches



Stability of North Anchors in 4 Pro Batches



Stability of 4 North Anchors across 6 Batches



DIF: Background

1. **Definition (Camilli & Shepard, 1995; Seo, Taherbhai, & Frantz, 2016)**

   Researchers' concern on possible item/test bias for/against particular groups.

   DIF refers to a phenomenon where the probability of successfully answering a specific item may differ from group to group, even after controlling the primary ability/trait that the test is designed to measure.

1. **Methodology (Hambleton, Swaminathan, & Rogers, 1991; Linacre, 2015; Thissen, Steinberg & Wainer, 1993)**

   Two Approached to detect DIF along with manifest variables (e.g., first language):

   1) Parametric approach: item response theory (IRT) –based chi-square test and IRT-based likelihood ratio test method
   2) Non-parametric approach: Mantel-Haenszel method and logistical regression method

   One of the most important pre-requisites is to obtain a common metric over groups: Equal-mean-difficulty (EMD), all-other-item (AOI), and constant-item (CI)

DIF: Methodology

3. **Psychometric Method for Current Study (Linacre, 2015; Pearson, 2015)**

   Rasch Rating Scale

   - Each item has its own item difficulty
   - The same step parameters across all the items

   Software: WINSTEPS 3.90

   Estimation: Joint Maximum Likelihood

   - Extreme score set to 0.5
   - Missing responses treated as missing

   Method to Obtain Common Matric:

   - The mean of item difficulties for each group was set to zero.
   - When the mean of item difficulties is set at zero, the two groups have equal mean item difficulty, and the impact of differences in the latent trait between groups is eliminated.
   - The two groups are considered to be put on a common metric.

   Critical Value for DIF:

   - Absolute value = 0.6

Current Data for Analysis

4. **Data**

   DIF analysis was performed on IRT2 explained above.

   6,938 raters (442 experts and 6,496 online respondents) who responded to 2,001 descriptors were analyzed for the present study.

GSE Descriptors:

Descriptor Set

- Academic (AL), General (GL), Professional (PL), Young Learner (YL). Plus North anchors (NO) and Council of Europe descriptors (CE).

Skill

- Listening, Reading, Writing, and Speaking

Raters:

Background Variables Used for DIF:

- Age-group that a teacher teaches
- First language that a teacher speaks
- Continent where a teacher lives

Reference vs. Focal Group:

Reference Group: English, Population

Focal Group: Spanish, Each Subgroup (e.g., AL-Group; Asia-Group)

Results (1): Descriptive Statistics-Raters

Table 1.1
Age-Group That He or She Teaches

Age-Group	Frequency	Percent
GL	3,097	51.96
PL	1,288	21.61
AL	1,012	16.98
YL	563	9.45
Total	5,960	100.00

Results (2): Descriptive Statistics-Raters

Table 1.2
First Language That He or She Speaks

Language	Frequency	Percent
English	1,554	58.07
Spanish	1,122	41.93
Total	2,676	100.00

Results (3): Descriptive Statistics-Raters

Table 1.3

Continent Where He or She Lives

Continent	Frequency	Percent
North America	444	8.02
South America	1,100	19.86
Europe	3,646	65.84
Africa	11	0.20
Asia	291	5.25
Australia/New Zealand	46	0.83
Total	5,538	100.00

Note. Four dominant countries for Asia: China, Japan, Korea, and Vietnam

Results (4): Descriptive Statistics-Raters

Table 1.4
Awareness of CEFR

CEFR	Frequency	Percent
Detailed	1,779	26.10
General	4,706	69.03
Heard	248	3.64
Never	84	1.23
Total	6,817	100.00

Results (4): Descriptive Statistics-Raters

Table 1.5
Years of Teaching

Years of Teaching	Frequency	Percent
Over 5	6,101	89.48
2 to 5	647	9.49
Less than 2	70	1.03
Total	6,818	100.00

Results (5): Descriptive Statistics-Descriptors

Table 1.6

Descriptor Set of GSE

Descriptor Set	Frequency	Percent
AL	484	26.20
CE	87	4.71
GL	482	26.10
NO	60	3.25
PL	403	21.82
YL	331	17.92
Total	1,847	100.00

Results (6): Descriptive Statistics-Descriptors

Table 1.7
Skill of GSE Descriptors

Skill	Frequency	Percent
Listening	308	17.10
Reading	356	19.77
Speaking	600	33.31
Writing	537	29.82
Total	1,801	100.00

Results (7): Age-Group DIF

Table 2.1
Age-Group by Descriptor Set

Descriptor Set	Age-Group That He or She Teaches							
	GL		PL		AL		YL	
	DIF No	DIF Yes	DIF No	DIF Yes	DIF No	DIF Yes	DIF No	DIF Yes
AL	431	52	396	87	385	0	**261**	**124**
	89%	11%	82%	18%	100%	0%	**68%**	**32%**
CE	87	0	87	0	85	2	**53**	**34**
	100%	0%	100%	0%	98%	2%	**61%**	**39%**
GA	436	41	395	81	397	79	**247**	**121**
	91%	9%	8298%	17%	83%	17%	**67%**	**33%**
NO	55	3	54	4	53	5	**35**	**19**
	95%	5%	93%	7%	91%	9%	**65%**	**35%**
PL	403	0	403	0	391	12	**182**	**131**
	100%	0%	100%	0%	97%	3%	**58%**	**42%**
YL	**113**	**93**	**60**	**44**	**47**	**57**	**331**	**0**
	55%	**45%**	**58%**	**42%**	**45%**	**55%**	**100%**	**0%**
Total	1,525	189	1,395	216	1,358	155	1,109	429
	89%	11%	87%	13%	90%	10%	**72%**	**28%**

Results (8): Age-Group DIF

Table 2.2.
Age-Group DIF by Skill

Skill	Age Group That He or She Teaches							
	GL		PL		AL		YL	
	DIF No	DIF Yes	DIF No	DIF Yes	DIF No	DIF Yes	DIF No	DIF Yes
Listening	235	41	211	43	208	24	**199**	**64**
	85%	15%	83%	17%	90%	10%	**76%**	**24%**
Reading	281	42	260	44	248	33	**232**	**74**
	87%	13%	86%	14%	88%	12%	**76%**	**24%**
Speaking	516	49	476	56	470	48	**353**	**145**
	91%	9%	89%	11%	91%	9%	**71%**	**29%**
Writing	447	57	402	73	387	49	**298**	**127**
	89%	11%	85%	15%	89%	11%	**70%**	**30%**
Total	1479	189	1349	216	1313	154	**1,082**	**410**
	89%	11%	86%	14%	90%	11%	**73%**	**27%**

Results (9): First Language DIF

Table 3.1
First Language DIF by Descriptor Set

Descriptor Set	Total Frequency	N-Count		Percent	
		DIF N	DIF Y	DIF N	DIF Y
AL	385	275	110	71.43	28.57
CE	87	75	12	86.21	13.79
GL	387	301	86	77.78	22.22
NO	54	45	9	83.33	16.67
PL	403	296	107	73.45	26.55
YL	**331**	**188**	**143**	**56.80**	**43.20**
Total	1,647	1,194	453		

Results (10): First Language DIF

Table 3.2
First Language DIF by Skill

Skill	Total Frequency	N-Count		Percent	
		DIF N	DIF Y	DIF N	DIF Y
Listening	267	198	69	74.16	25.84
Reading	**319**	**221**	**98**	**69.28**	**30.72**
Speaking	**558**	**386**	**172**	**69.18**	**30.82**
Writing	457	336	121	73.52	26.48
Total	1,601	1,141	460		

Results (11): First Language DIF (in YL)

Table 3.3
First Language DIF in YL

Skill	Total Frequency	N-Count		Percent	
		DIF No	DIF Yes	DIF No	DIF Yes
Listening	68	41	27	60.29	39.71
Reading	84	48	36	57.14	42.86
Speaking	**94**	**45**	**49**	**47.87**	**52.13**
Writing	85	54	31	63.53	36.47
Total	331	189	142		

Results (12): Continent-Group DIF

Table 4.1
Continent DIF by Descriptor Set

Descriptor Set	Continent							
	North America		South America		Europe		Asia	
	DIF No	DIF Yes	DIF No	DIF Yes	DIF No	DIF Yes	DIF No	DIF Yes
AL	410	74	424	60	474	10	**307**	**177**
	85%	15%	88%	12%	98%	207%	**63%**	**37%**
CE	78	9	84	3	87	0	**66**	**21**
	90%	10%	97%	3%	100%	0%	**76%**	**24%**
GL	409	73	409	73	477	5	**331**	**151**
	85%	15%	85%	15%	99%	104%	**69%**	**31%**
NO	48	12	55	5	59	1	**43**	**17**
	80%	20%	92%	8%	98%	2%	**72%**	**28%**
PL	355	48	359	44	401	2	**254**	**132**
	88%	12%	89%	11%	100%	1%	**66%**	**34%**
YL	**203**	**128**	267	64	327	4	**222**	**109**
	61%	**39%**	81%	19%	99%	1%	**67%**	**33%**
Total	1,503	344	1,598	249	1,825	22	**1,223**	**607**
	81%	19%	87%	13%	99%	1%	**67%**	**33%**

Results (13): Continent-Group DIF

Table 4.2
Continent DIF by Skill

Skill	Continent							
	North America		South America		Europe		Asia	
	DIF No	DIF Yes	DIF No	DIF Yes	DIF No	DIF Yes	DIF No	DIF Yes
Listening	243	65	273	35	306	2	**205**	**103**
	79%	21%	89%	11%	99%	1%	**67%**	**33%**
Reading	289	67	293	63	348	8	**239**	**114**
	81%	19%	82%	18%	98%	2%	**68%**	**32%**
Speaking	502	98	515	85	593	7	**399**	**195**
	84%	16%	86%	14%	99%	1%	**67%**	**33%**
Writing	427	110	472	65	532	5	**345**	**184**
	80%	20%	88%	12%	99%	1%	**65%**	**35%**
Total	1,461	340	1,553	248	1,779	22	**1,188**	**596**
	81%	19%	86%	14%	99%	1%	**67%**	**33%**

Results (14): Continent-Group DIF

Table 4.3
Continent DIF on Skill by North American vs. Asian Groups: AL

Skill	Adult Learner			
	North America		Asia	
	DIF No	DIF Yes	DIF No	DIF Yes
Listening	96	21	72	45
	82%	18%	62%	38%
Reading	103	15	82	36
	87%	13%	69%	31%
Speaking	96	12	72	36
	89%	11%	67%	33%
Writing	115	26	81	60
	82%	18%	**57%**	**43%**
Total	410	74	307	177
	85%	15%	63%	37%

Results (15): Continent-Group DIF

Table 4.4
Continent DIF on Skill by North American vs. Asian Groups: YL

Skill	Young Learner			
	North America		Asia	
	DIF No	DIF Yes	DIF No	DIF Yes
Listening	39	29	43	25
	57%	**43%**	63%	37%
Reading	58	26	55	29
	69%	31%	65%	35%
Speaking	59	35	62	32
	63%	37%	66%	34%
Writing	47	38	62	23
	55%	**45%**	**73%**	**27%**
Total	203	128	222	109
	61%	39%	67%	33%

Results (16): Continent-Group DIF

Table 4.5
Continent DIF on Skil by North American vs. Asian Groups: GL

Skill	General Learner			
	North America		Asia	
	DIF No	DIF Yes	DIF No	DIF Yes
Listening	74	9	63	20
	89%	11%	76%	24%
Reading	70	15	60	25
	82%	**18%**	71%	29%
Speaking	130	23	103	50
	85%	15%	**67%**	**33%**
Writing	135	26	105	56
	84%	16%	**65%**	**35%**
Total	409	73	331	151
	85%	15%	69%	31%

Results (17): Continent-Group DIF

Table 4.6
Continent DIF on Skill by North American vs. Asian Groups: PL

Skill	Professional Learner			
	North America		Asia	
	DIF No	DIF Yes	DIF No	DIF Yes
Listening	23	3	18	8
	88%	12%	69%	31%
Reading	56	11	40	24
	84%	**16%**	**63%**	**38%**
Speaking	174	20	122	66
	90%	10%	**65%**	**35%**
Writing	102	14	74	34
	88%	12%	69%	31%
Total	355	48	254	132
	88%	12%	66%	34%

Discussion

1. **Methodological perspective (Linacre, 2015; Pearson, 2015):**
   a. Rasch Rating Scale model still fits well with many missing responses
   b. Absolute value (i.e., 0.6) for DIF works.
2. **Empirical DIF evidence (Eckes, 2011; Taherbhai, Seo, & O'Malley, 2014):**
   a. YL age-group
   b. First language-group
   c. Asia continent-group: bigger sample size for subgroup analysis?
3. **Practical perspective (Seo, Taherbhai, & Frantz, 2016):**
   a. Feedback for writer/content staff training

References

1. Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: SAGE Publications.
2. Eckes, T. (2011) *Introduction to many-facet Rash measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang GmbH.
3. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications, Inc.
4. Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
5. Linacre, J. M. (2015). *A user's guide to WINSTEPS: Rasch-model computer programs*, Chicago, IL. Available: http://www.winsteps.com.
6. Pearson Assessment and Information. (2015, December). *The Australian National Assessment Program – Literacy and Numeracy (NAPLAN) 2015 Technical Report*. San Antonio, TX: Author.
7. Seo, D., Taherbhai, H., & Frantz, R. (2016, in press). Psychometric Evaluation and Discussions of English Language Learners' Listening Comprehension. *International Journal of Listening.*
8. Taherbhai, H., Seo, D., & O'Malley. (2014). Formative information using student growth percentiles for the quantification of English language learners' progress in language acquisition. *Applied Measurement in Education, 27,* 196-213
9. Taherbhai, H. Seo, D., & Bowman, T. (2012). Comparison of paper–pencil and online performances of students with learning disabilities, *British Educational Research Journal, 38:1,* 61-74.
10. Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Vocabulary

GSE Vocabulary - Summary

- **A graded lexical inventory of general English for adults which complements the functional guidance found in the CEFR** (Council of Europe, 2001) descriptors

- It identifies **what meanings learners are expected to produce at different proficiency levels on the CEFR/GSE**

- It classifies vocabulary according to the CoE categorization in **Specific Notions** [*fork*], **General Notions** [*somewhere*], and **Functions** [*I am sorry*]

- It combines **L1 corpus frequency + communicative usefulness** – as per judgments provided by pool of teachers

- It was developed **to inform teaching, course design, and assessment**

GSE Vocabulary - Search tool

- **Searchable online** by keyword, part of speech, topic/subtopic, collocations, and proficiency level. Staging site available at http://gsevocab-beta.ayr-digital.co.uk/ **> demo**

- It includes:
  - **~ 28,000 lemmas**
  - **~ 36,000 word meanings**
  - **2,370 phrasal verbs**
  - **7,355 phrases**
  - **80,000+ collocations**

GSE Vocabulary - Scaling methodology

Combine frequency data and teacher ratings to produce a weighted algorithm to grade vocabulary

1. Statistical analysis and data cleaning
2. Theoretical assumptions based on research evidence
3. Data modelling

GSE Vocabulary - Scaling methodology

The first action is to check the soundness of the rating data

- **Measures per rater**

Overall average of a rater compared with other raters
Standard deviation of the raters' ratings
Correlation of a rater's ratings with average of all raters
One rater removed

- **Measures per word: 450,157 ratings gathered**

Remove rating if >1.5 distance from average rating
4.7% ratings removed
Percentage of ratings within two adjacent categories
words <70% in two adjacent categories: 1575 (=4.9%)
Correlation of ratings and frequency data
r= 0.44

GSE Vocabulary - Scaling methodology

Combine ratings and frequency data

- Frequency rank data ranging from 3 to 54,496
- Rescale to 1-5 scale
- Then combine average ratings with Freq Rank
- Weight rating data for their reliability

Rating average = Ra

Reliability of rating data = r_{Rating}

- $Ra \times r_{Rating} + F_{rank\ m} \times (1- r_{Rating})$

GSE Vocabulary - Scaling methodology

- Grabe (2008), chapter 13

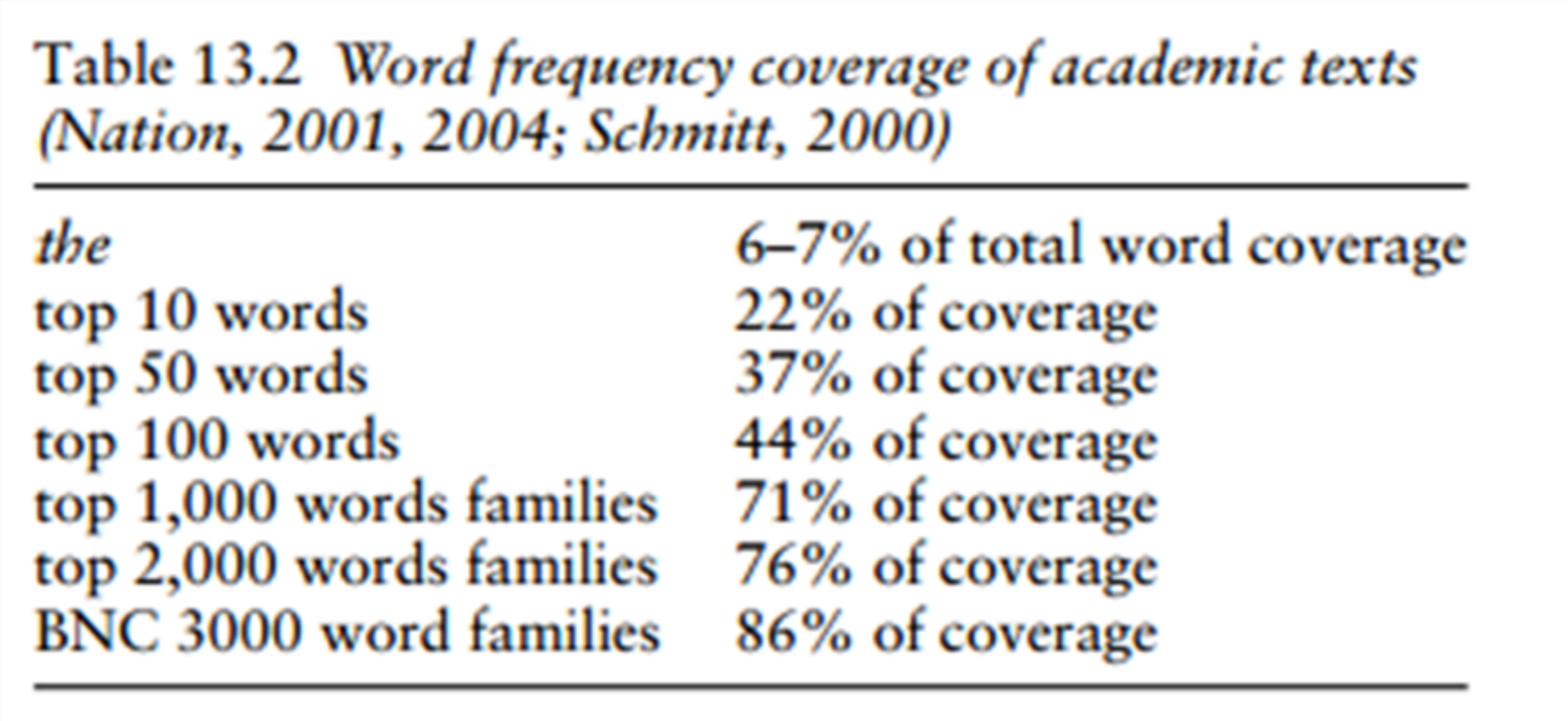
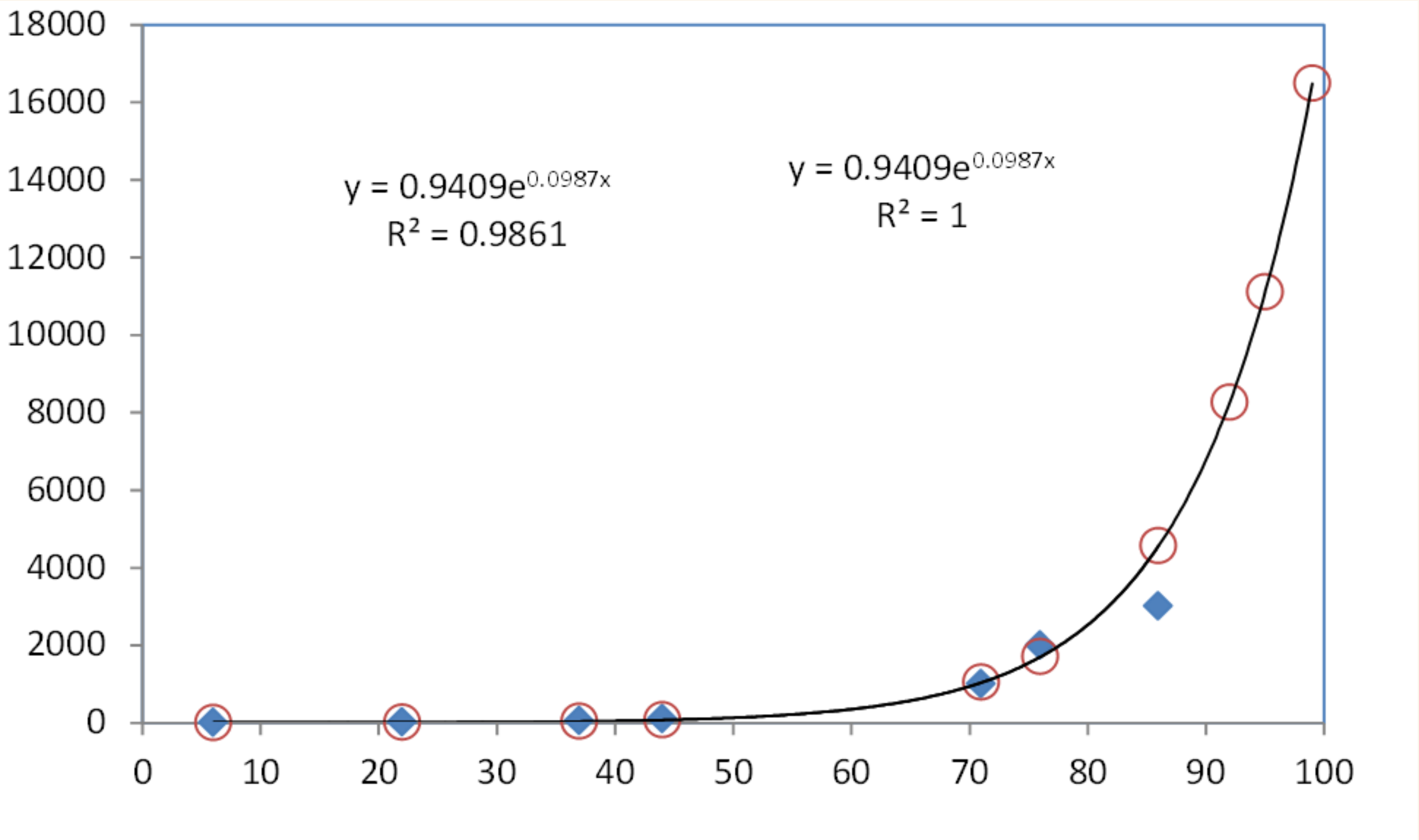


Table 13.2 *Word frequency coverage of academic texts (Nation, 2001, 2004; Schmitt, 2000)*

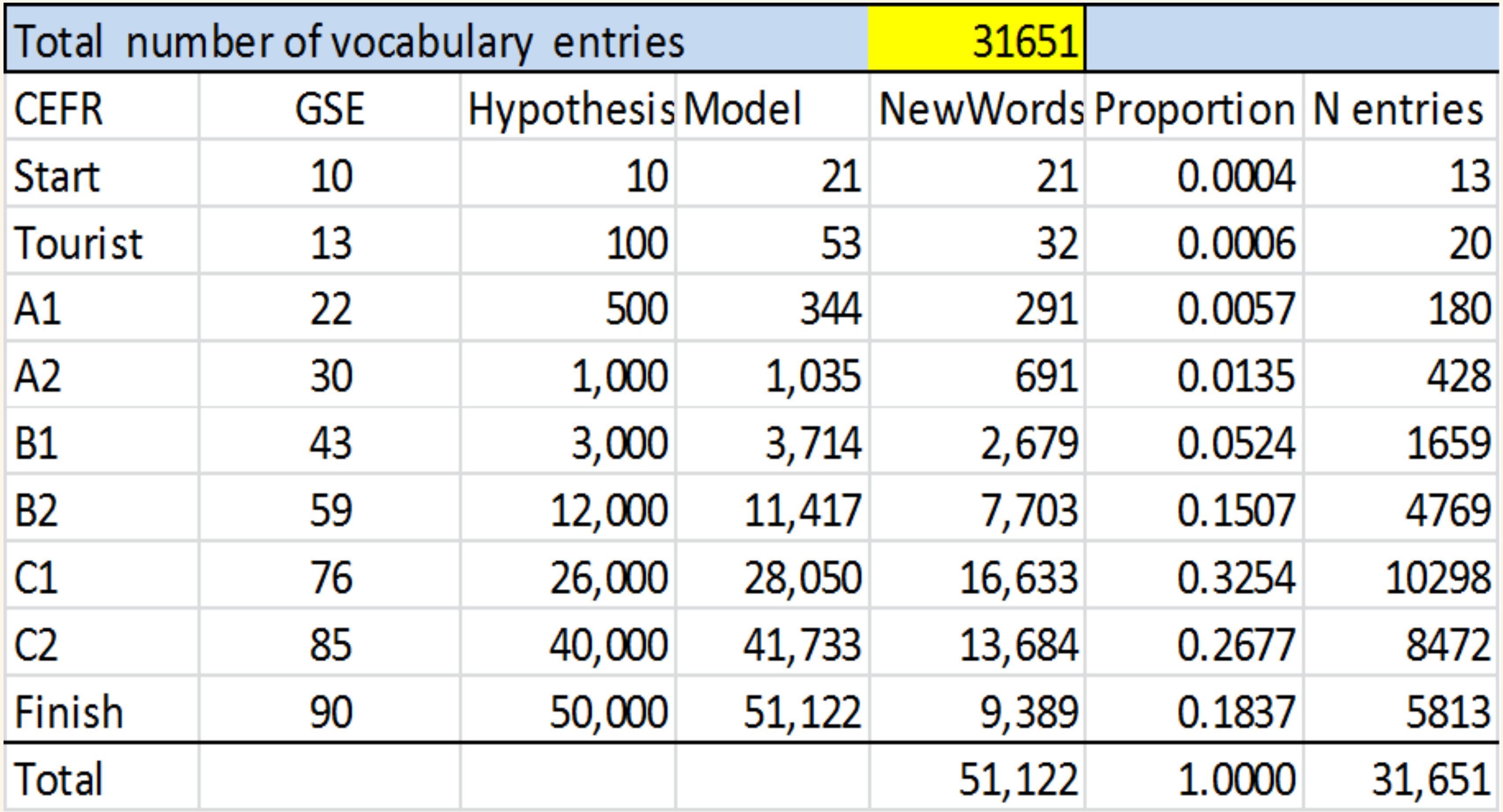
the	6–7% of total word coverage
top 10 words	22% of coverage
top 50 words	37% of coverage
top 100 words	44% of coverage
top 1,000 words families	71% of coverage
top 2,000 words families	76% of coverage
BNC 3000 word families	86% of coverage

GSE Vocabulary - Scaling methodology



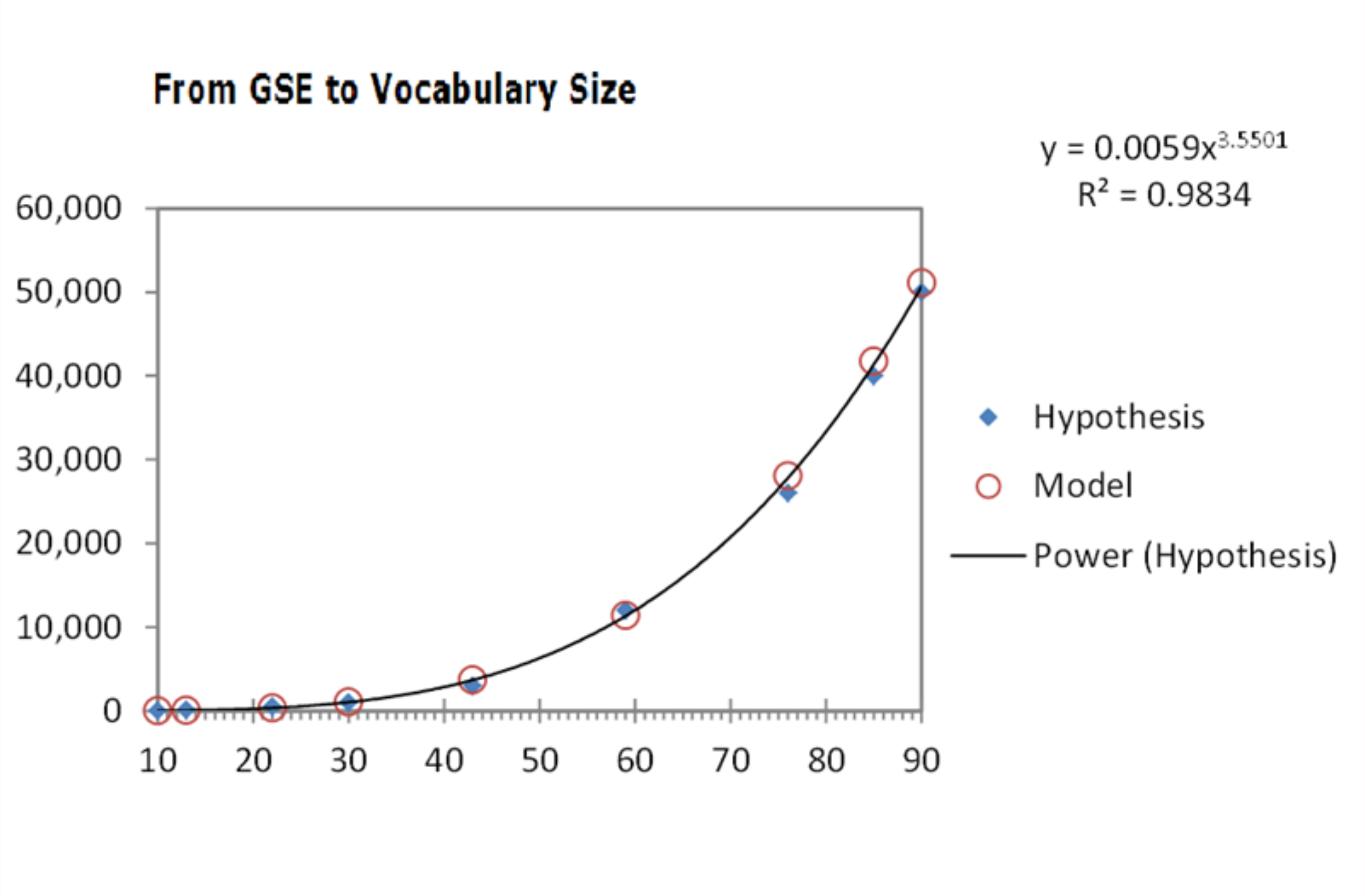
GSE Vocabulary - Scaling methodology

- **10k lemmas** for academic studies (Hazenberg & Hulstijn, 1996)

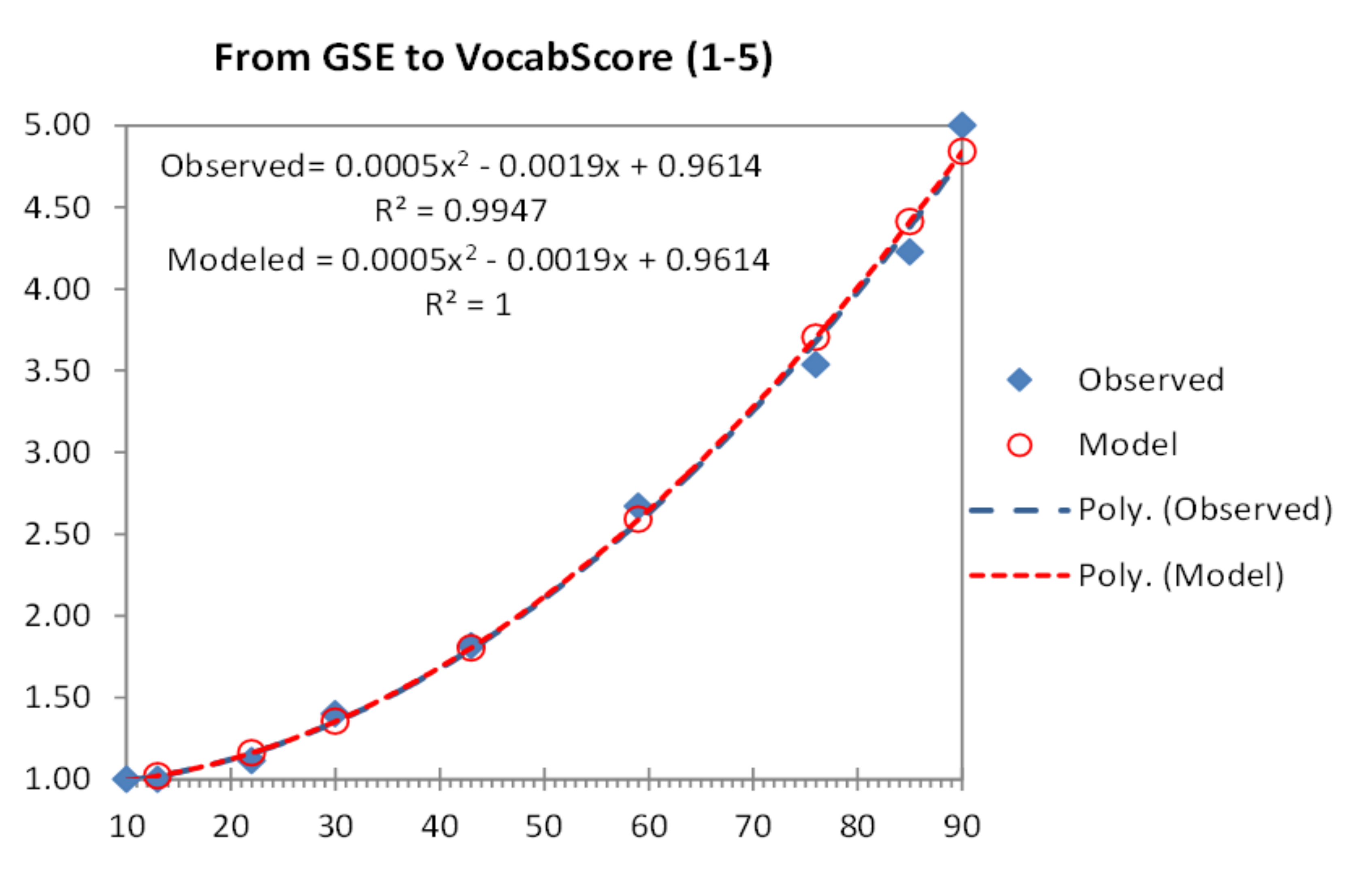


Total number of vocabulary entries				31651		
CEFR	GSE	Hypothesis	Model	NewWords	Proportion	N entries
Start	10	10	21	21	0.0004	13
Tourist	13	100	53	32	0.0006	20
A1	22	500	344	291	0.0057	180
A2	30	1,000	1,035	691	0.0135	428
B1	43	3,000	3,714	2,679	0.0524	1659
B2	59	12,000	11,417	7,703	0.1507	4769
C1	76	26,000	28,050	16,633	0.3254	10298
C2	85	40,000	41,733	13,684	0.2677	8472
Finish	90	50,000	51,122	9,389	0.1837	5813
Total				51,122	1.0000	31,651

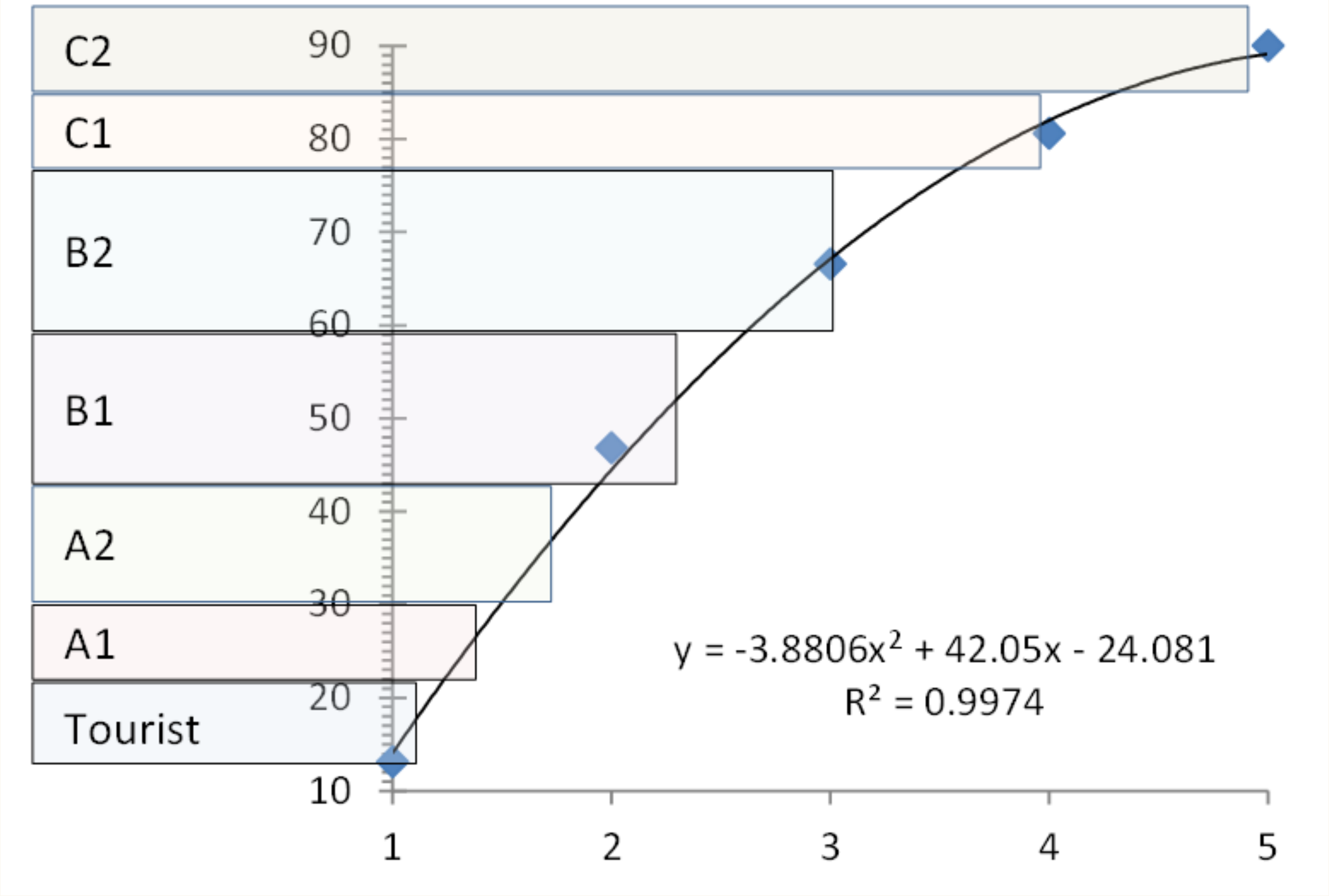
GSE Vocabulary - Scaling methodology



GSE Vocabulary - Scaling methodology



GSE Vocabulary - Scaling methodology


C2
C1
B2
B1
A2
A1
Tourist
y = -3.8806x² + 42.05x - 24.081
R² = 0.9974