Developing an efficient EAP placement test using integrated tasks to assess receptive and productive skills

LARC 2024

Friday, April 12th

Rebecca Yeager and Alfonso Martinez

University Of Iowa

Integrated Assessment for Placement

- Integrated Assessment (IA) outperforms independent tasks:
  - Predictive validity (Llosa & Malone, 2019)
  - Cognitive validity (Rukthong, 2021)
- However, IA is underused for local placement purposes due to two proposed difficulties:

1. impracticality (Weigle, 2004)
2. trouble identifying support needs for receptive skills (Asención, 2008)

1. Test Design for Efficiency

- Diagnosing the problem:
  - IA raters lose time locating source information, assessing quality of source use, and identifying citation mechanics (Gebril & Plakans, 2014)
- Treating the problem:
  - 1. Since lexical overlap from listening sources is associated with summary quality (Kyle, 2020), the rubric explicitly allows patchwriting from listening sources
  - 2. Rather than checking every borrowed idea for accuracy, raters only hold students accountable for accurately representing five key ideas from each source, which raters assess using a checklist (Park & Yan, 2019)
  - 3. Since all students must eventually take a Rhetoric class which teaches citation conventions, citations are not required

2. Test Design for Receptive Skills

Three-pronged approach:

1. Analytic rubric with multiple sub-scores for receptive and productive skills (Ohta et al., 2018; Shin & Ewert, 2015)

2. Multiple integrated tasks (Asención, 2008)

3. A few selected-response tasks targeting details, inference, and vocabulary skills (Rukthong, 2021)

Placement Test Design

- Written Test:

Input	Output
❖ 2 reading passages ❖ 1 listening passage	❖3 written summaries + 1 argumentative writing task ❖10 sentence identification questions for reading ❖5 MCQs for reading and 5 MCQs for listening (vocab, detail, and inference)

- Oral interview

- **Rating:**
  - 2 human raters + 3rd if 2+ bands apart
  - Analytic rubric with 7 sub-scores and 5 bands each

Task 1+4	Task 2+4	Task 3+4	Task 4	All 4 Tasks		
Source 1 Rep	Source 2 Rep	Source 3 Rep	Argument	Coherence	Vocabulary	Grammar

Test Performance: Rasch Analysis

Sp23: Rater Severity M: 0.00, SD: .24

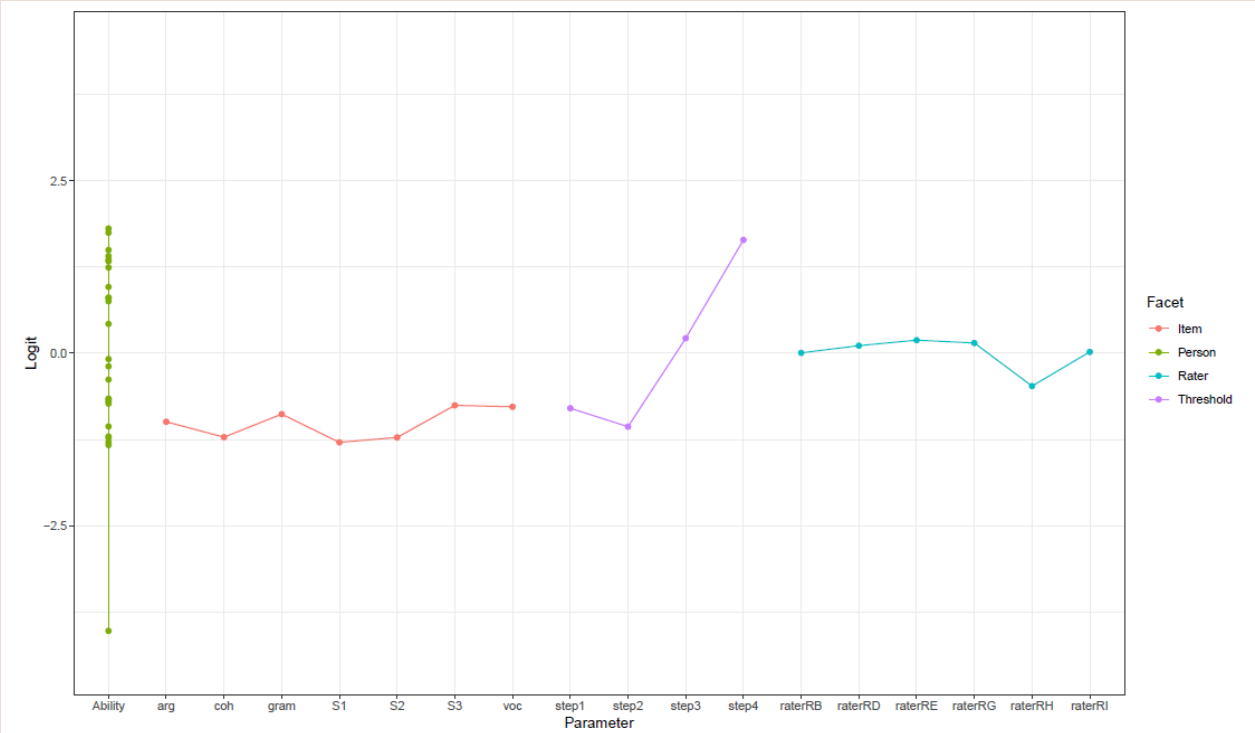
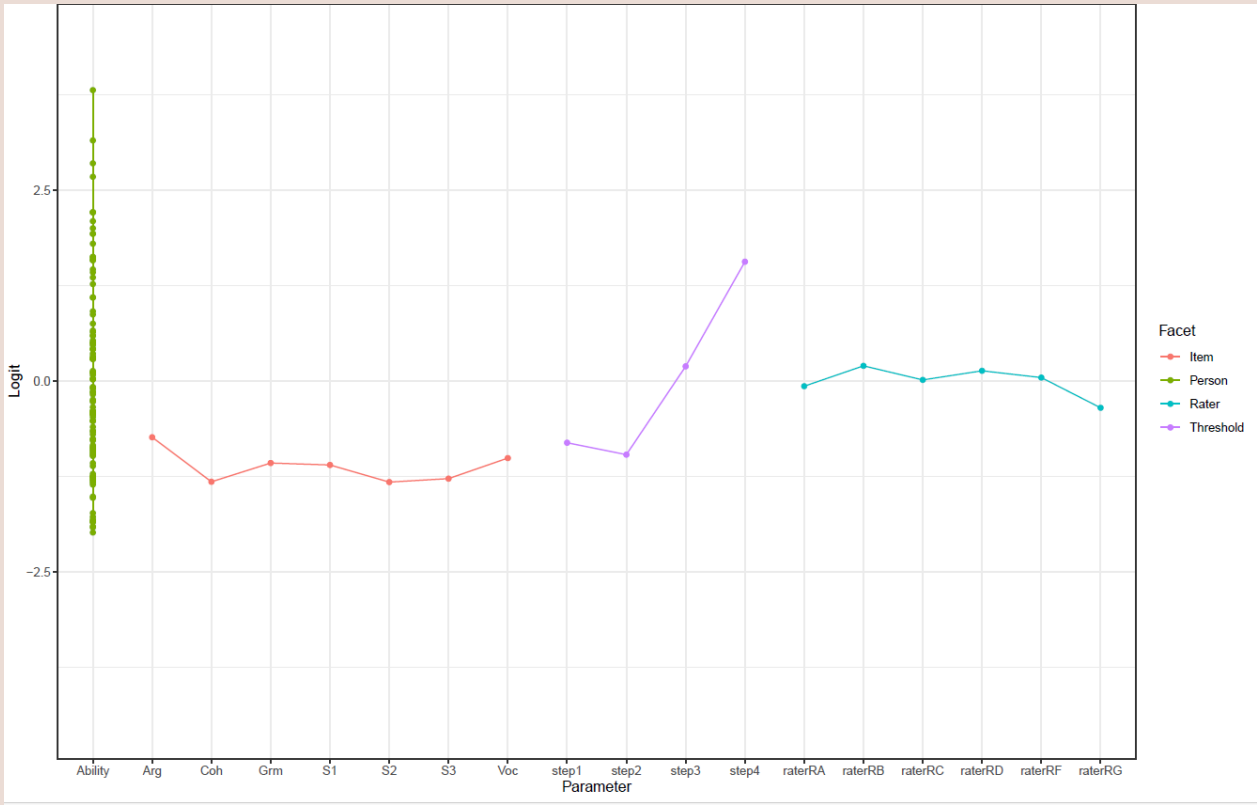
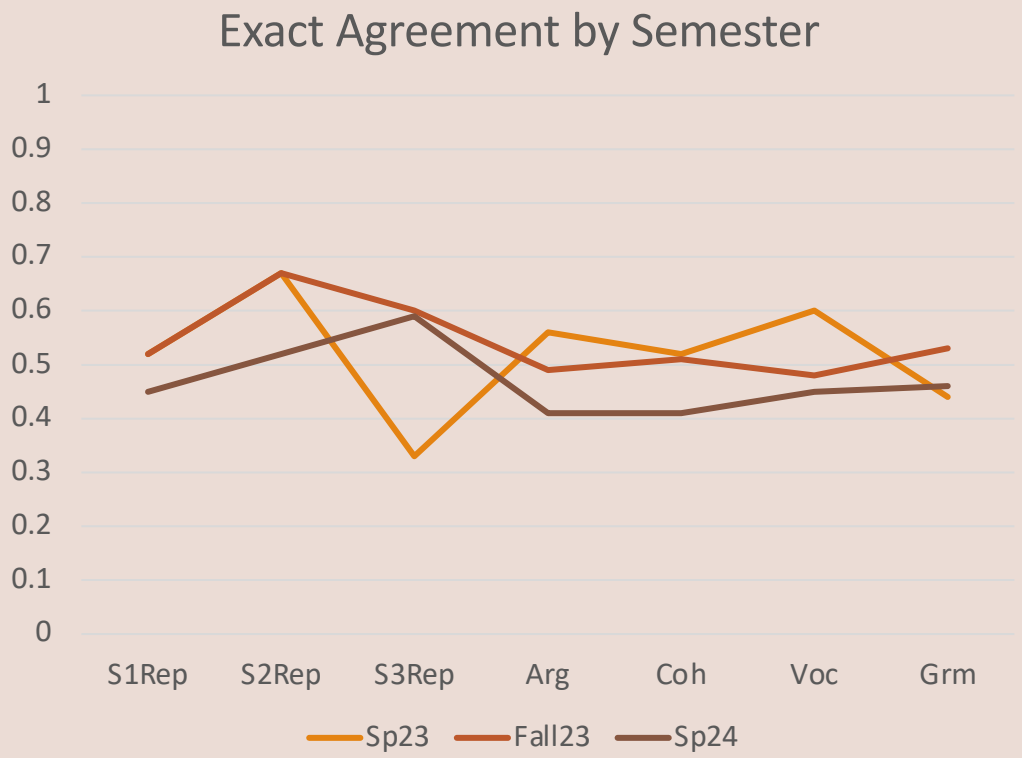


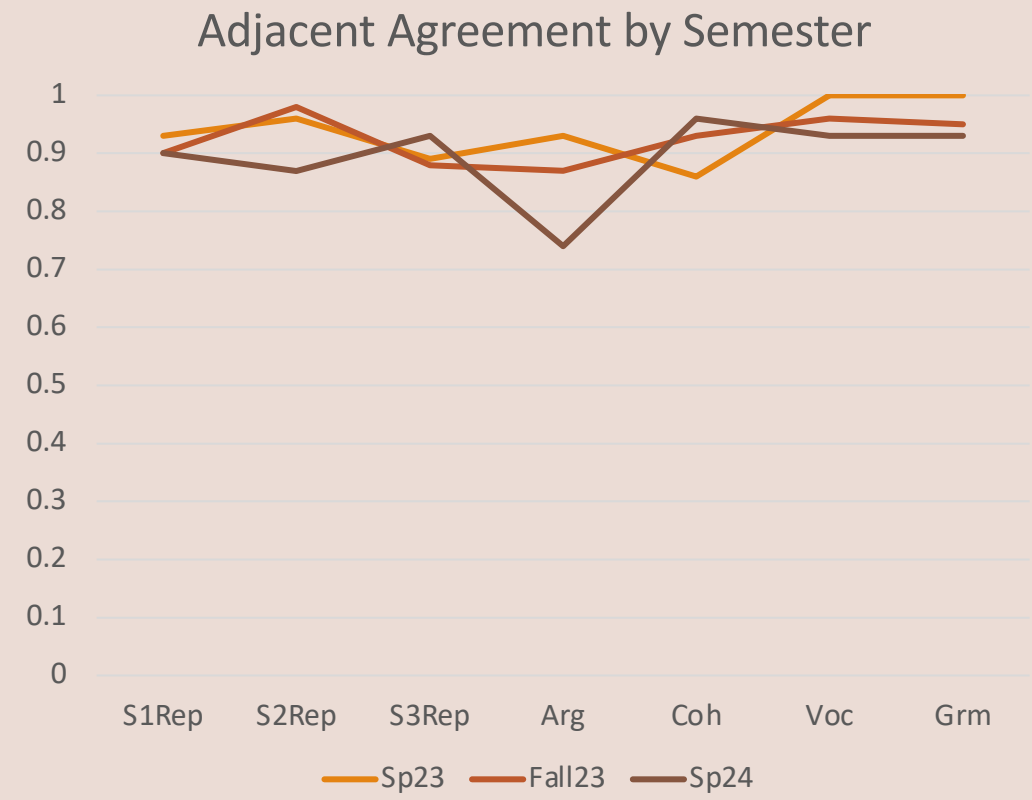
Figure 7: Variable Map Showing Location of Each Facet Element From Rasch Analysis

Fall23: Rater Severity M: 0.00, SD: .19



Test Performance: IRR





Test Performance: Efficiency

Semester	Mean rating time-to-decision including 1st, 2nd, and occasional 3rd raters
Sp23	30.43 minutes per sample
Fall23	23 minutes per sample
Sp24	no data (rating remote and asynchronous because of polar vortex, campus closure, and power outage)

Test Performance: Receptive Skills

- Diagnostic intake checklists
  - All instructors fill out checklists about each student at the end of the second week of classes
  - 5 items target key learning outcomes
  - Students who have already mastered learning outcomes are waived from class requirement

Semester	Reading	Listening	Writing	Oral Skills	Total New Student Misplacements
Sp23		1			1/100 = 1%
Fall23			3		3/472 = .63%
Sp24	2	1		1	4/100 = 4%

- Receptive skills and productive skills have equally low misplacement rates

References

- Asención, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7*, 140–150. DOI: 10.1016/j.jeap.2008.04.001
- Gebril, A., and Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing, 21*: 56-73. DOI: 10.1016/j.asw.2014.03.002
- Hall, J., and Walker-Cecil, K. (2015). Interdepartmental communication: The key to ensuring international student preparedness. Presentation at MIDTESOL 2015.
- Kyle, K. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing, 45,* advance online publication. DOI: 10.1016/j.asw.2020.100467
- Llosa, L., and Malone, M. (2019). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. *Language Testing, 36*(2), 235-263. DOI: 10.1177/0265532218763456
- Ohta, R., Plakans, L, and Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing, 38,* 21-36. DOI: 10.1016/j.asw.2018.08.001
- Park, H., and Yan, X. (2019). An investigation into rater performance with a holistic scale and a binary, analytic scale on an ESL writing placement test. *Papers in Language Testing and Assessment, 8*(2): 34-64. Retrieved from https://experts.illinois.edu/en/publications/an-investigation-into-rater-performance-with-a-holistic-scale-and
- Rukthong, A. (2021). MC listening questions vs. integrated listening-to-summarize tasks: What listening abilities do they assess? *System, 97,* advance online publication. DOI: 10.1016/j.system.2020.102439
- Shin, S., and Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing, 32*(2), 259-281. DOI: 10.1177/0265532214560257
- Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*: 27-55. DOI: 10.1016/j.asw.2004.01.002

OVERVIEW

- Integrated Assessment for placement
- Our Test Design
  1. How Our Test Targets Efficiency
  2. How Our Test Targets Receptive Skills
- Test Performance
- Background slides:
  - Test Development

Test Development

- Needs analysis
  - Faculty survey (Hall and Walker-Cecil, 2015)
  - Rhetoric survey on changes to syllabi and tasks post-COVID
  - Sampled syllabi and textbooks from most popular freshman classes
- Reviewed literature on integrated assessment
- Looked at examples of published integrated assessments
  - IELTS
  - TOEFL
- Drafted specifications, reading and listening materials, items, and rubric
- Piloted in Fall 2021
  - Rater training session 1
- Revised and piloted large scale in Spring 2022
  - Rater training session 2
- Fall 2022: ready for use

Task 1 and 4	Task 2 and 4	Task 3 and 4	Task 4	All 4 Tasks		
Source 1 Representation	**Source 2 Representation**	**Source 3 Representation**	**Argumentation**	**Coherence**	**Vocabulary**	**Grammar**
0-1 key ideas from Source 1 are represented accurately in student's words.	0-1 key ideas from Source 2 are represented accurately in student's words.	0-1 key ideas from Source 3 are represented accurately.	Essay fails to follow instructions about topic or source use, OR opinion, reasons, and support are incomprehensible.	Relative importance of ideas (opinion>reasons>support) is impossible to distinguish, OR connections between ideas and sentences are not attempted.	Vocabulary is extremely limited and often incomprehensible, OR the only appropriate academic vocabulary comes from sources.	Essay consists of single words, short phrases, and simple sentences which are often incomprehensible, OR the only comprehensible language comes from sources.
Two key ideas from Source 1 are represented accurately in student's words.	Two key ideas from Source 2 are represented accurately in student's words.	Two key ideas from Source 3 are represented accurately.	Essay compares two sources but fails to state an opinion, states opinion without reasons, OR support is contradictory, irrelevant, insufficient, vague, or unclear.	Relative importance of ideas (opinion>reasons>support) is difficult to distinguish, OR connections between ideas and sentences often break down due to missing, misapplied, or vague cohesive devices.	Vocabulary is basic, repetitive, awkward, and frequently confusing, OR most appropriate academic vocabulary comes from sources.	Essay consists of simple sentences, attempts complex sentences which are incomprehensible, OR directly copies language from sources with few attempts to change the grammar.
Three key ideas from Source 1 are represented accurately in student's words.	Three key ideas from Source 2 are represented accurately in student's words.	Three key ideas from Source 3 are represented accurately.	Essay states an opinion comparing two sources but with only one reason, OR support is sometimes contradictory, irrelevant, insufficient, vague, or unclear.	Relative importance of ideas (opinion>reasons>support) is indicated inconsistently, OR connections between ideas and sentences sometimes break down due to missing, misapplied, or vague cohesive devices.	Vocabulary that is basic, repetitive, awkward, or unclear is more common than vocabulary that is precise, varied, and clear, OR sometimes over-uses vocabulary from sources.	Essay uses a variety of complex structures, verbs, and word forms with inconsistent clarity, OR attempts to paraphrase language from sources but doesn't change enough grammatically.
Four key ideas from Source 1 are represented accurately in student's words.	Four key ideas from Source 2 are represented accurately in student's words.	Four key ideas from Source 3 are represented accurately.	Essay states an opinion comparing two sources with two or more reasons, BUT support is slightly contradictory, irrelevant, insufficient, vague, or unclear.	Relative importance of ideas (opinion>reasons>support) is mostly clear, OR connections between ideas and sentences may be slightly unclear due to missing, misapplied, or vague cohesive devices.	Vocabulary is generally precise, varied, and clear, with occasional basic, repetitive, awkward, or unclear expressions, OR slightly over-uses words from sources	Essay uses a variety of complex structures, verbs, and word forms with general clarity, OR successfully paraphrases most language from sources except for a few short phrases.
All five key ideas from Source 1 are represented accurately in student's words.	All five key ideas from Source 2 are represented accurately in student's words.	All five key ideas from Source 3 are represented accurately.	Essay states an opinion comparing two sources with two or more reasons, AND support is consistent, relevant, sufficient, specific, and clear.	Relative importance of ideas (opinion>reasons>support) is clearly indicated, AND connections between ideas and sentences are clearly indicated with sophisticated and appropriate cohesive devices.	Vocabulary is consistently precise, varied, and clear, AND only borrows necessary words from sources.	Essay skillfully incorporates a variety of complex structures, verbs, and word forms, AND successfully paraphrases all language from sources.

	Reading	Listening	Writing	Oral Skills
Task Subscores *(Passing subscores must be 3.5 or higher)*	Source 1 Representation: ___/5	Source 3 Representation: ___/5	Argumentation: ___/5	Pronunciation: ___/5
	Source 2 Representation: ___/5	Listening Questions: ___/5	Coherence: ___/5	Fluency: ___/5
	Sentence Identification: ___/10 ÷ 2 = ___/5	Oral Listening: ___/5	Vocabulary: ___/5	Oral Grammar: ___/5
	Reading Questions: ___/5		Written Grammar: ___/5	
Subscore Pass?	Y / N	Y / N	Y / N	Y / N
Total by Skill *(Passing totals must be 80% or higher)*	___/20	___/15	___/20	___/15
Total Pass?	Y / N	Y / N	Y / N	Y / N
Placement Decision *(Circle One)*	Pass --- No Pass	Pass --- No Pass	Pass --- No Pass	Pass --- No Pass

Test Performance

- Sp23 Rasch Analysis Estimates

Parameter	Facet	Estimate	Standard Error
arg	Item	-0.99	0.18
coh	Item	-1.22	0.19
gram	Item	-0.88	0.18
S1	Item	-1.29	0.19
S2	Item	-1.22	0.19
S3	Item	-0.75	0.17
voc	Item	-0.78	0.18
Threshold 1	Threshold	-0.80	0.12
Threshold 2	Threshold	-1.06	0.12
Threshold 3	Threshold	0.22	0.12
Threshold 4	Threshold	1.64	0.21
Rater RB	Rater	0.01	0.11
Rater RD	Rater	0.11	0.12
Rater RE	Rater	0.19	0.13
Rater RG	Rater	0.15	0.11
Rater RH	Rater	-0.48	0.14
Rater RI	Rater	0.02	0.27

Table 1: Estimated Item Parameters and Standard Errors From Rasch Analysis

Test Performance

- Fall23 Rasch Analysis Estimates

Parameter	Facet	Estimate	Standard Error
Arg	Item	-0.73	0.07
Coh	Item	-1.32	0.08
Gram	Item	-1.07	0.08
S1	Item	-1.10	0.08
S2	Item	-1.32	0.08
S3	Item	-1.27	0.08
Voc	Item	-1.01	0.08
Threshold 1	Threshold	-0.80	0.06
Threshold 2	Threshold	-0.96	0.05
Threshold 3	Threshold	0.20	0.05
Threshold 4	Threshold	1.57	0.09
Rater RA	Rater	-0.06	0.06
Rater RB	Rater	0.20	0.04
Rater RC	Rater	0.02	0.06
Rater RD	Rater	0.14	0.06
Rater RF	Rater	0.05	0.07
Rater RG	Rater	-0.35	0.13

Test Performance