Randomly Distributed

Comparative Judgement

An alternative approach to essay grading

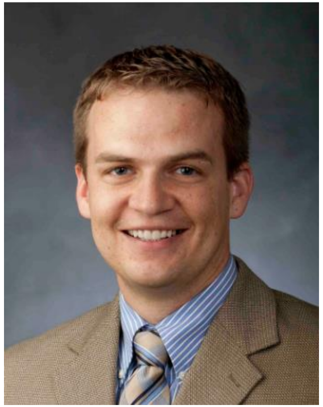
MEET THE research team

Mornie Sims

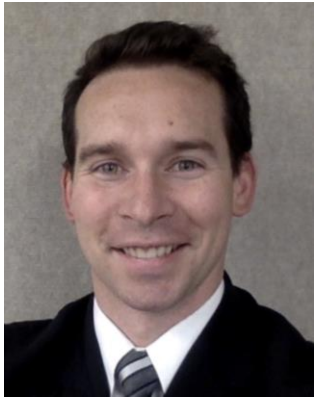
Dr. Cox

Dr. Eckstein

Dr.
Hartshorn



Judson Hart



Dr. Wilcox

Reliability
consistency

Cold War

Validity
authenticity

reliability?

1880s – inconsistent scoring

⬇ reliability → ? validity

indirect → MC testing

component skills

highly reliable

strongly correlated with writing grades

validity?

1961 Study – opposite effect

spurious correlations (# of bathrooms)

teacher focus on component skills (Braddock, et al.)

writing → active skill

MC → passive, undue attention to less important features

direct

RELIABILITY IN

writing assessment

Rubrics

Training

Double-rating

Adjudication

MFRM

THE rubric METHOD

- Absolute judgment
- External standard
- Training/calibration

RANDOMLY DISTRIBUTED

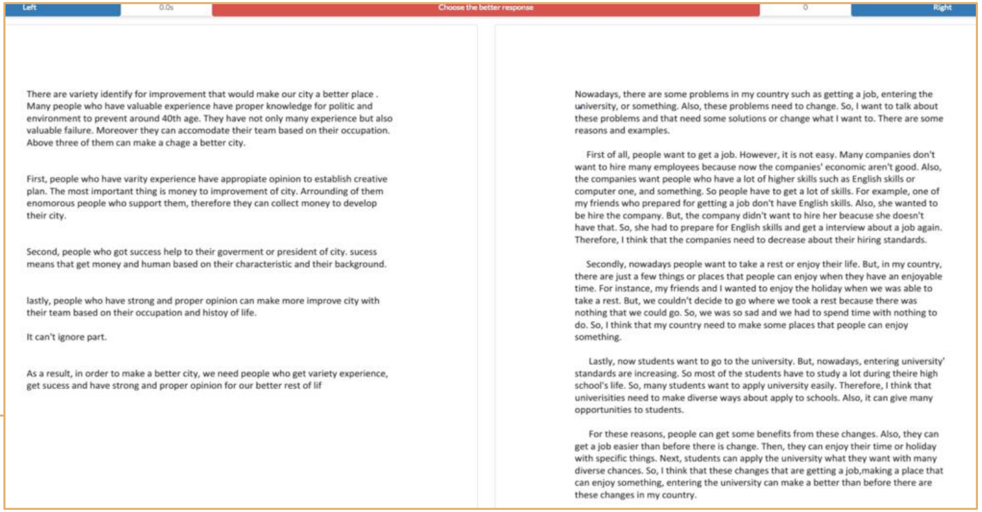
comparative judgment

- Comparison
- Relative choice
- Instinctual skill

> “There is no absolute judgment. All judgments are comparisons of one thing to another.”
>
> [Donald Laming]

RR	&	RDCJ
Explicit comparison		Implicit comparison
Minimizes training		Training for consensus
Minimizes bias		Unavoidable bias
Inherent algorithm		MFRM

HOW IT works.



There are variety identify for improvement that would make our city a better place . Many people who have valuable experience have proper knowledge for politic and environment to prevent around 40th age. They have not only many experience but also valuable failure. Moreover they can accomodate their team based on their occupation. Above three of them can make a chage a better city.

First, people who have varity experience have appropiate opinion to establish creative plan. The most important thing is money to improvement of city. Arrounding of them enomorous people who support them, therefore they can collect money to develop their city.

Second, people who got success help to their goverment or president of city. sucess means that get money and human based on their characteristic and their background.

lastly, people who have strong and proper opinion can make more improve city with their team based on their occupation and histoy of life.

It can't ignore part.

As a result, in order to make a better city, we need people who get variety experience, get sucess and have strong and proper opinion for our better rest of lif

Nowadays, there are some problems in my country such as getting a job, entering the university, or something. Also, these problems need to change. So, I want to talk about these problems and that need some solutions or change what I want to. There are some reasons and examples.

First of all, people want to get a job. However, it is not easy. Many companies don't want to hire many employees because now the companies' economic aren't good. Also, the companies want people who have a lot of higher skills such as English skills or computer one, and something. So people have to get a lot of skills. For example, one of my friends who prepared for getting a job don't have English skills. Also, she wanted to be hire the company. But, the company didn't want to hire her beacuse she doesn't have that. So, she had to prepare for English skills and get a interview about a job again. Therefore, I think that the companies need to decrease about their hiring standards.

Secondly, nowadays people want to take a rest or enjoy their life. But, in my country, there are just a few things or places that people can enjoy when they have an enjoyable time. For instance, my friends and I wanted to enjoy the holiday when we was able to take a rest. But, we couldn't decide to go where we took a rest because there was nothing that we could go. So, we was so sad and we had to spend time with nothing to do. So, I think that my country need to make some places that people can enjoy something.

Lastly, now students want to go to the university. But, nowadays, entering university' standards are increasing. So most of the students have to study a lot during theire high school's life. So, many students want to apply university easily. Therefore, I think that univerisities need to make diverse ways about apply to schools. Also, it can give many opportunities to students.

For these reasons, people can get some benefits from these changes. Also, they can get a job easier than before there is change. Then, they can enjoy their time or holiday with specific things. Next, students can apply the university what they want with many diverse chances. So, I think that these changes that are getting a job,making a place that can enjoy something, entering the university can make a better than before there are these changes in my country.

demo

nomoremarking.com

https://www.nomoremarking.com/demo1

test it!

nomoremarking.com

https://www.nomoremarking.com/judges/reg/sLRRwmGAe65Wx3mbv

CJ RATIONALE

Steedle and Ferrara, 2016

CJ eliminates common scoring biases

- Strictness vs leniency
- Central or extreme tendencies

Additionally

- it is less cognitively demanding/time consuming per judgment
- it requires less training
- evidence suggests that it is highly accurate *(Gill & Bramley, 2008)*

comparative judgment

…is a promising alternative, BUT is it…

Reliable and Practical?

and

Can we trust the results?

research question

How does traditional rubric rating compare with MFRM (many facet Rasch model) and RDCJ (randomly distributed Comparative Judgment) in an ESL setting in terms of **reliability**, **validity**, and **practicality**?

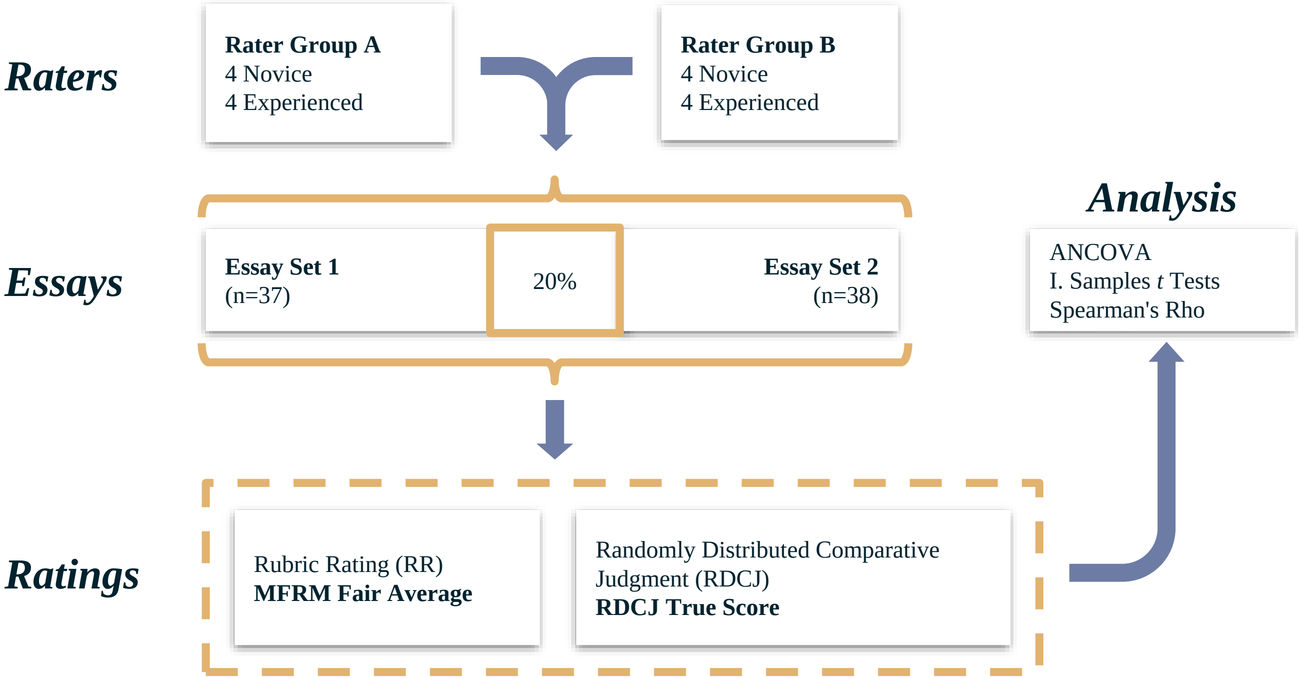


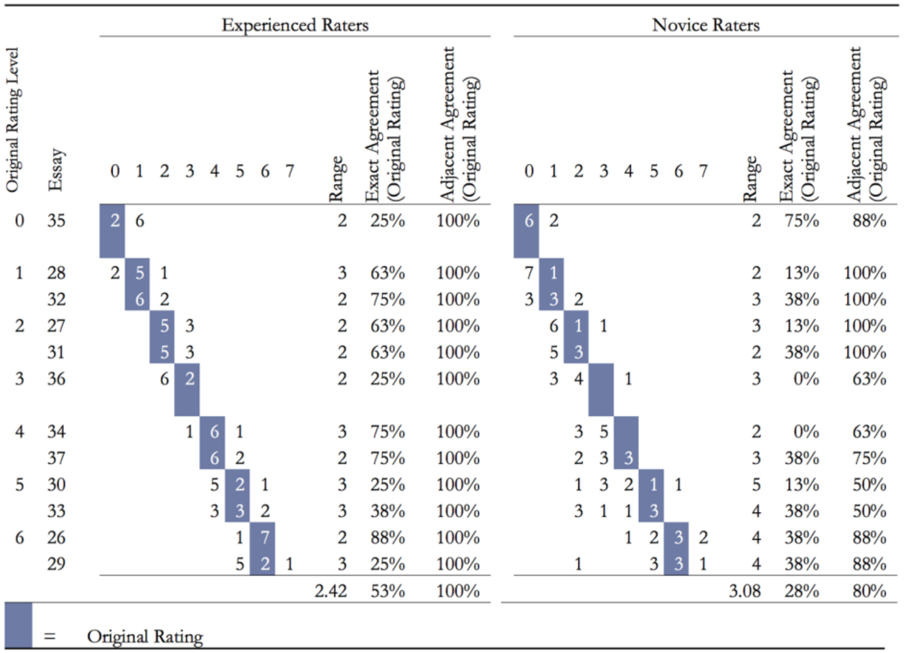
Figure 2. Study design to compare traditional rubric rating (RR) to multi-facet Rasch modeling (MFRM) and randomly distributed comparative judgment (RDCJ). Analysis of variance (ANOVA) run to test for effects on rating time and Spearman's rho used to correlate between MFRM adjusted fair average, the study rubric rating fair averages, and RDCJ true scores to show evidence of validity.

SELECTED Essay

	Essay Rating Levels						
Languages	0	1	2	3	4	5	6
Arabic	-	-	-	-	-	-	1
Chinese	-	1	2	-	-	-	-
French	-	-	-	-	1	1	-
Japanese	-	-	1	1	-	-	-
Korean	-	3	1	1	-	-	1
Mongolian	-	1	-	1	-	-	-
Portuguese	-	-	1	1	1	2	2
Russian	-	-	-	-	1	-	-
Spanish	2	3	5	5	5	6	6
Thai	-	-	-	-	2	1	-
Turkish	-	1	-	-	-	-	-
Totals	2	9	10	9	10	10	10

Rubric Rating

WITHOUT MFRM



		Experienced Raters											Novice Raters										
Original Rating Level	Essay	0	1	2	3	4	5	6	7	Range	Exact Agreement (Original Rating)	Adjacent Agreement (Original Rating)	0	1	2	3	4	5	6	7	Range	Exact Agreement (Original Rating)	Adjacent Agreement (Original Rating)
0	35	2	6							2	25%	100%	6	2							2	75%	88%
1	28	2	5	1						3	63%	100%	7	1							2	13%	100%
	32		6	2						2	75%	100%	3	3	2						3	38%	100%
2	27			5	3					2	63%	100%		6	1	1					3	13%	100%
	31			5	3					2	63%	100%		5	3						2	38%	100%
3	36			6	2					2	25%	100%		3	4		1				3	0%	63%
4	34				1	6	1			3	75%	100%			3	5					2	0%	63%
	37					6	2			2	75%	100%			2	3	3				3	38%	75%
5	30					5	2	1		3	25%	100%			1	3	2	1	1		5	13%	50%
	33					3	3	2		3	38%	100%			3	1	1	3			4	38%	50%
6	26						1	7		2	88%	100%					1	2	3	2	4	38%	88%
	29						5	2	1	3	25%	100%			1			3	3	1	4	38%	88%
										2.42	53%	100%									3.08	28%	80%

▨ = Original Rating

Evidence

RELIABILITY & VALIDITY

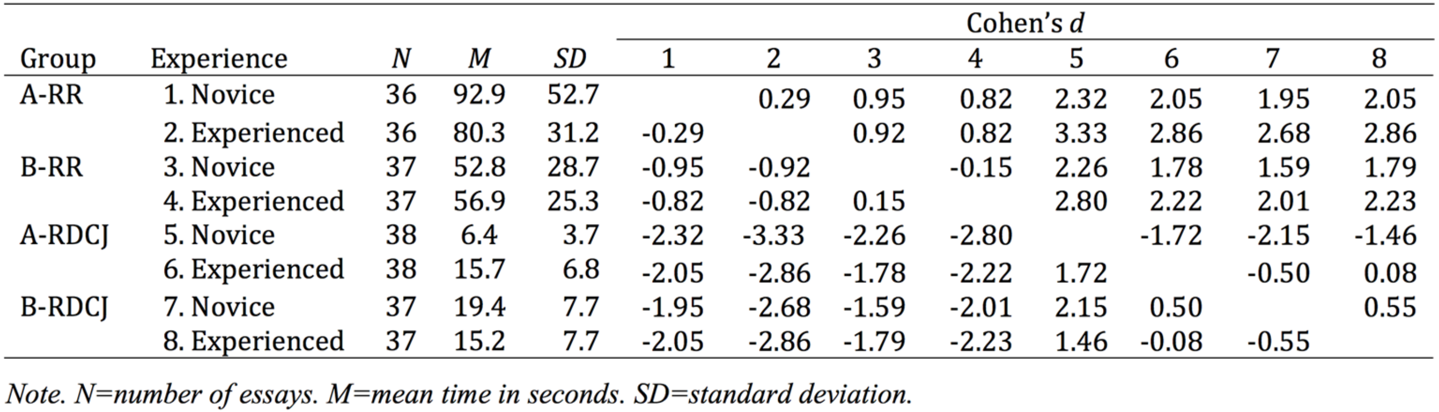
				Reliability	Validity
Group	Experience	Mode	*N*	Separation	rho
A	Novice	RR	36	0.96	0.94
		RDCJ	38	0.91	0.90
	Experienced	RR	36	0.98	0.95
		RDCJ	38	0.92	0.94
B	Novice	RR	37	0.96	0.96
		RDCJ	37	0.89	0.92
	Experienced	RR	37	0.96	0.94
		RDCJ	37	0.94	0.94

Note. RR=rubric rating; RDCJ=randomly distributed comparative judgment.

Practicality

DATA

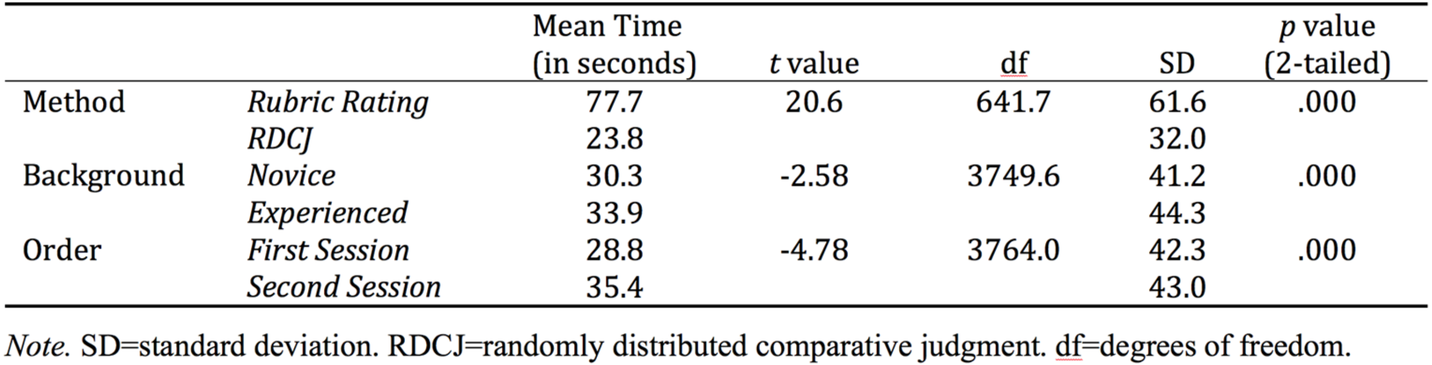
COHEN'S *d*



					Cohen's *d*							
Group	Experience	*N*	*M*	*SD*	1	2	3	4	5	6	7	8
A-RR	1. Novice	36	92.9	52.7		0.29	0.95	0.82	2.32	2.05	1.95	2.05
	2. Experienced	36	80.3	31.2	-0.29		0.92	0.82	3.33	2.86	2.68	2.86
B-RR	3. Novice	37	52.8	28.7	-0.95	-0.92		-0.15	2.26	1.78	1.59	1.79
	4. Experienced	37	56.9	25.3	-0.82	-0.82	0.15		2.80	2.22	2.01	2.23
A-RDCJ	5. Novice	38	6.4	3.7	-2.32	-3.33	-2.26	-2.80		-1.72	-2.15	-1.46
	6. Experienced	38	15.7	6.8	-2.05	-2.86	-1.78	-2.22	1.72		-0.50	0.08
B-RDCJ	7. Novice	37	19.4	7.7	-1.95	-2.68	-1.59	-2.01	2.15	0.50		0.55
	8. Experienced	37	15.2	7.7	-2.05	-2.86	-1.79	-2.23	1.46	-0.08	-0.55	

Note. N=number of essays. M=mean time in seconds. SD=standard deviation.

t TESTS

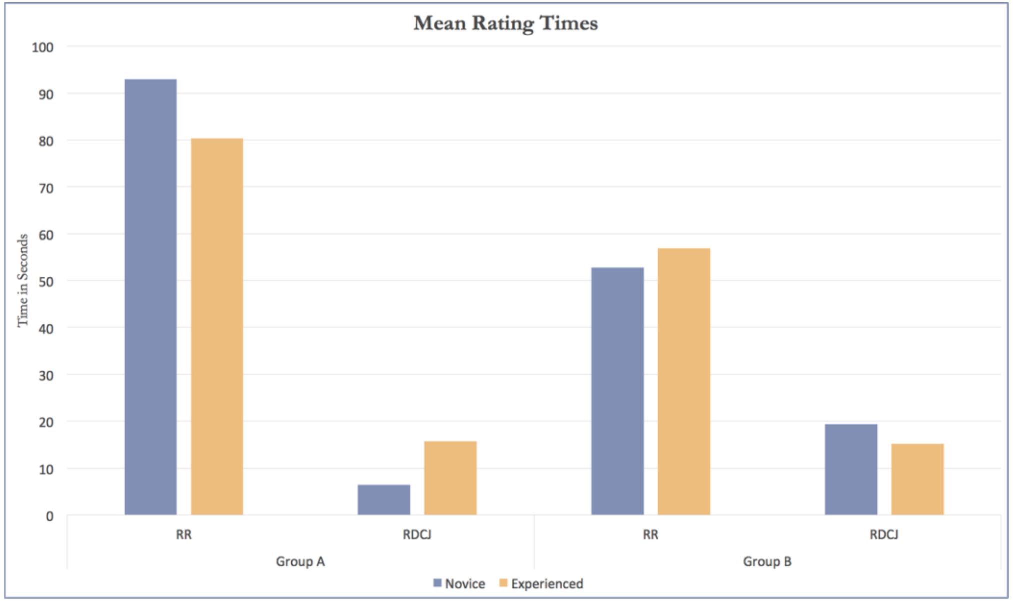


		Mean Time (in seconds)	*t* value	df	SD	*p* value (2-tailed)
Method	*Rubric Rating*	77.7	20.6	641.7	61.6	.000
	RDCJ	23.8			32.0	
Background	*Novice*	30.3	-2.58	3749.6	41.2	.000
	Experienced	33.9			44.3	
Order	*First Session*	28.8	-4.78	3764.0	42.3	.000
	Second Session	35.4			43.0	

Note. SD=standard deviation. RDCJ=randomly distributed comparative judgment. df=degrees of freedom.

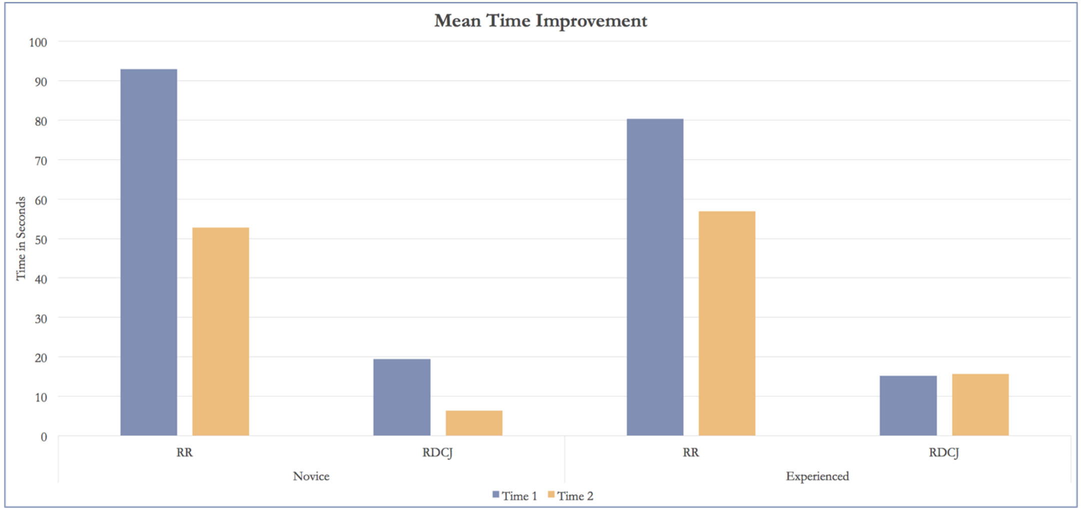
Covariance

ANALYSIS OF

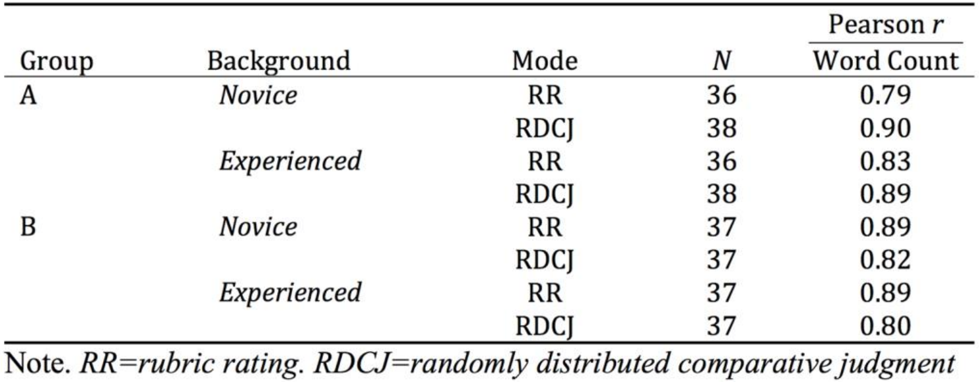


ANALYSIS OF

Covariance



LENGTH & RATINGS
essay



Group	Background	Mode	*N*	Pearson *r*
				Word Count
A	*Novice*	RR	36	0.79
		RDCJ	38	0.90
	Experienced	RR	36	0.83
		RDCJ	38	0.89
B	*Novice*	RR	37	0.89
		RDCJ	37	0.82
	Experienced	RR	37	0.89
		RDCJ	37	0.80

Note. *RR=rubric rating. RDCJ=randomly distributed comparative judgment*

CJ

APPLICATIONS

Barkhaoui, 2016

Bramley, 2015

Christodolou, 2016

Heldsinger & Humphrey, 2013

- Especially suited to productive tasks
  - Portfolios, essays, short answer
- Many subject areas
  - English, ESL, History, Geography
- Interesting Applications
  - Mathematical problem solving
  - Peer Assessment (highly reliable & correlated with expert ratings)

SUBJECT Areas

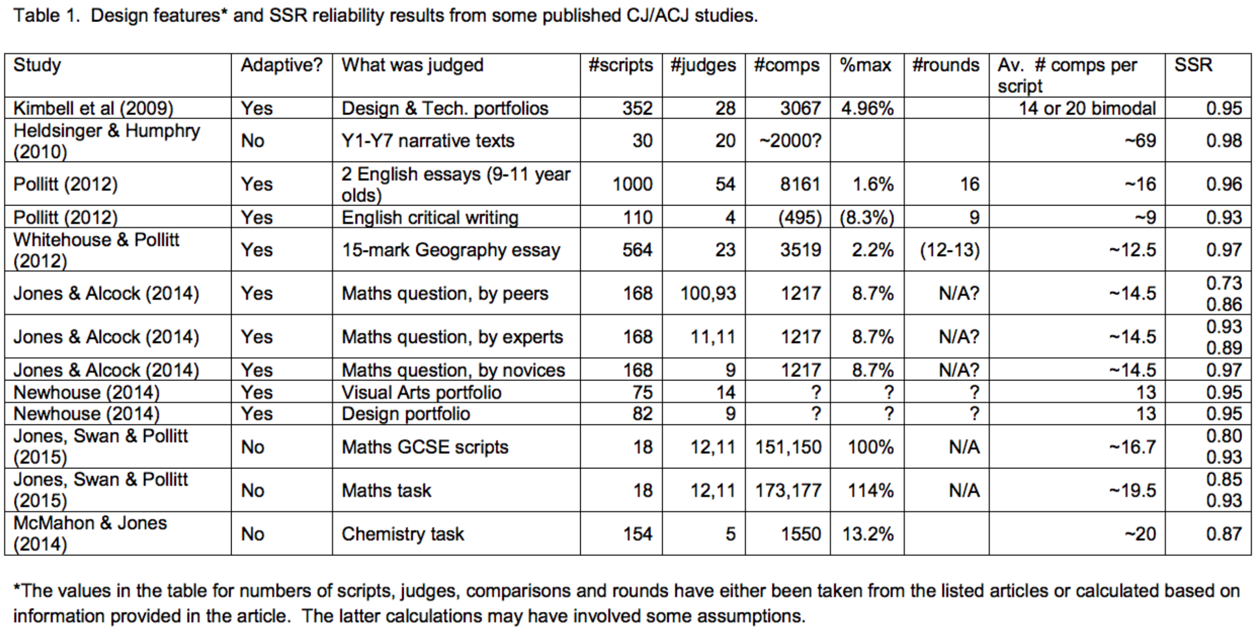
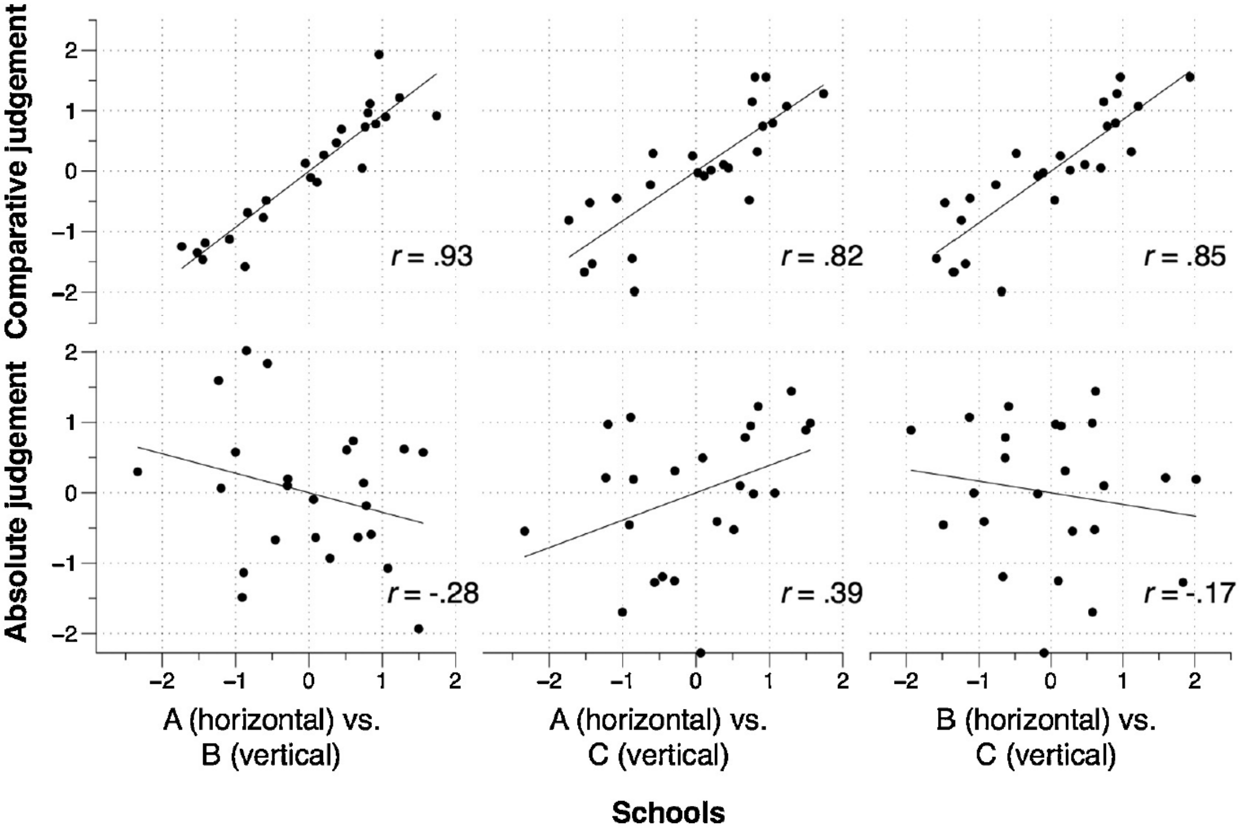


Table 1. Design features* and SSR reliability results from some published CJ/ACJ studies.

Study	Adaptive?	What was judged	#scripts	#judges	#comps	%max	#rounds	Av. # comps per script	SSR
Kimbell et al (2009)	Yes	Design & Tech. portfolios	352	28	3067	4.96%		14 or 20 bimodal	0.95
Heldsinger & Humphry (2010)	No	Y1-Y7 narrative texts	30	20	~2000?			~69	0.98
Pollitt (2012)	Yes	2 English essays (9-11 year olds)	1000	54	8161	1.6%	16	~16	0.96
Pollitt (2012)	Yes	English critical writing	110	4	(495)	(8.3%)	9	~9	0.93
Whitehouse & Pollitt (2012)	Yes	15-mark Geography essay	564	23	3519	2.2%	(12-13)	~12.5	0.97
Jones & Alcock (2014)	Yes	Maths question, by peers	168	100,93	1217	8.7%	N/A?	~14.5	0.73 0.86
Jones & Alcock (2014)	Yes	Maths question, by experts	168	11,11	1217	8.7%	N/A?	~14.5	0.93 0.89
Jones & Alcock (2014)	Yes	Maths question, by novices	168	9	1217	8.7%	N/A?	~14.5	0.97
Newhouse (2014)	Yes	Visual Arts portfolio	75	14	?	?	?	13	0.95
Newhouse (2014)	Yes	Design portfolio	82	9	?	?	?	13	0.95
Jones, Swan & Pollitt (2015)	No	Maths GCSE scripts	18	12,11	151,150	100%	N/A	~16.7	0.80 0.93
Jones, Swan & Pollitt (2015)	No	Maths task	18	12,11	173,177	114%	N/A	~19.5	0.85 0.93
McMahon & Jones (2014)	No	Chemistry task	154	5	1550	13.2%		~20	0.87

*The values in the table for numbers of scripts, judges, comparisons and rounds have either been taken from the listed articles or calculated based on information provided in the article. The latter calculations may have involved some assumptions.

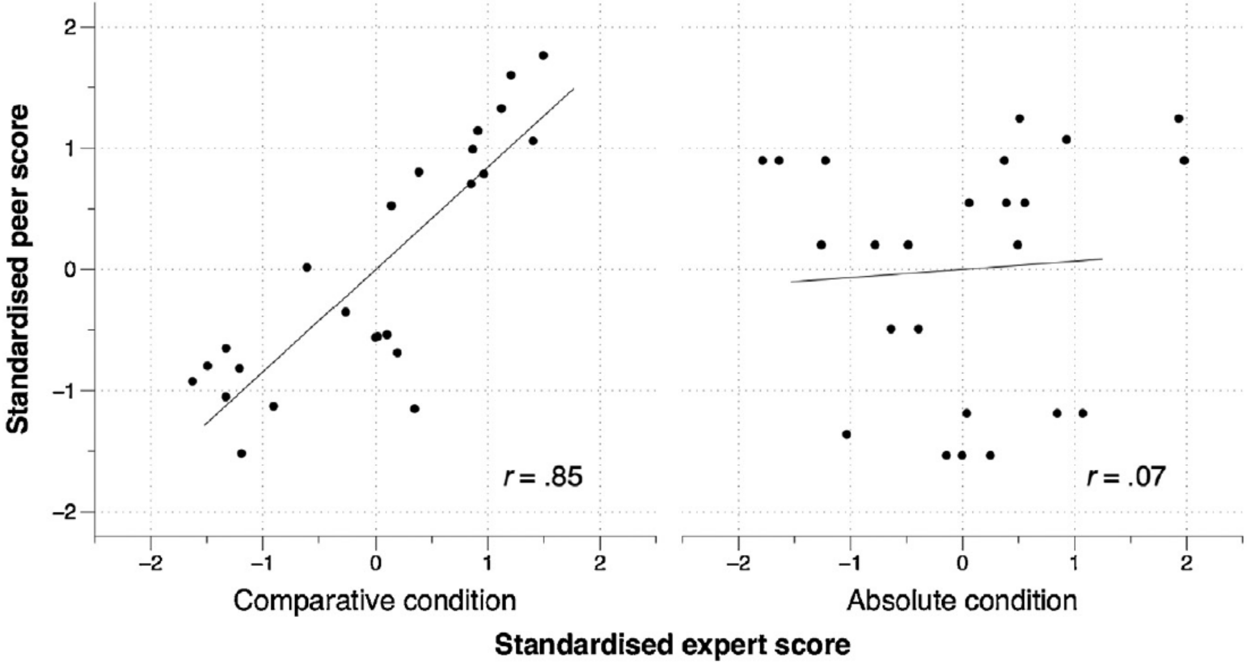
Peer
ASSESSMENT



Peer

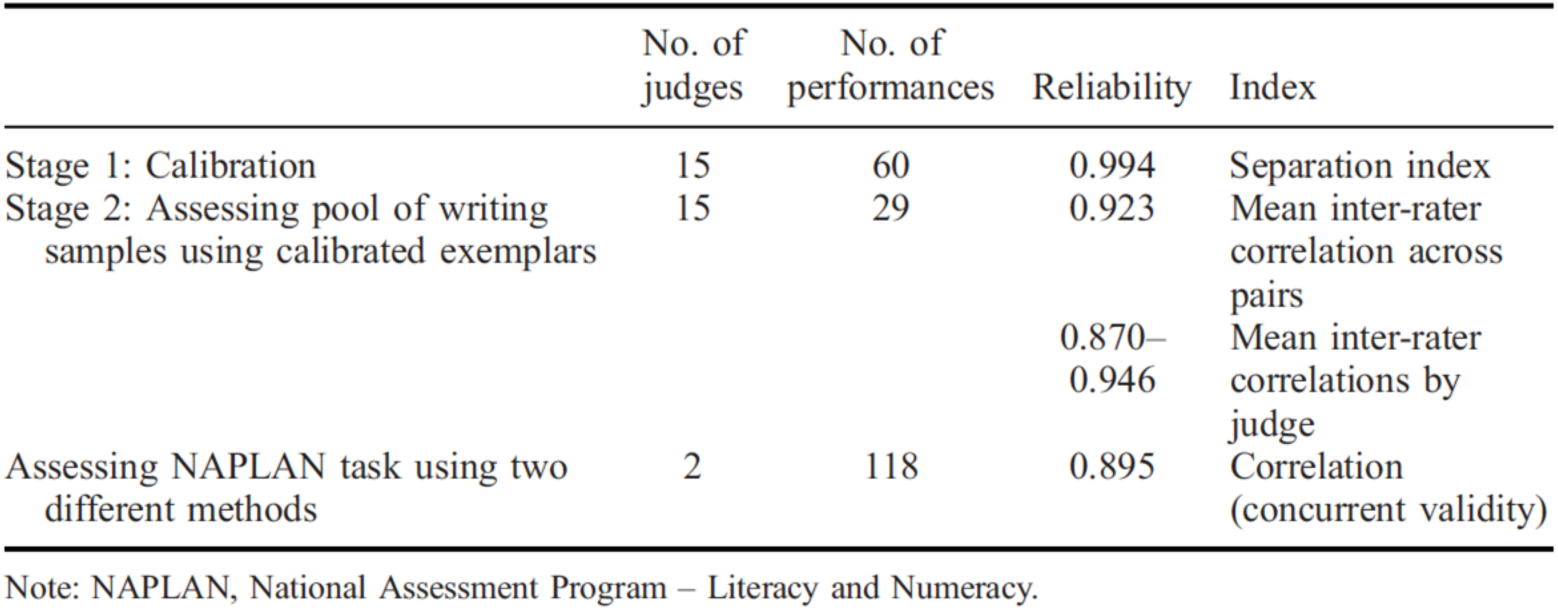
ASSESSMENT

(cont)



calibrated

EXEMPLARS



	No. of judges	No. of performances	Reliability	Index
Stage 1: Calibration	15	60	0.994	Separation index
Stage 2: Assessing pool of writing samples using calibrated exemplars	15	29	0.923	Mean inter-rater correlation across pairs
			0.870–0.946	Mean inter-rater correlations by judge
Assessing NAPLAN task using two different methods	2	118	0.895	Correlation (concurrent validity)

Note: NAPLAN, National Assessment Program – Literacy and Numeracy.

Comparative Judgment

thank you!

Mornie Sims
eslmornie@gmail.com

Dr. Troy Cox
Troy_cox@byu.edu

Dr. Matthew Wilcox
wilcoxmp@byu.edu

Dr. Grant Eckstein
grant_eckstein@byu.edu

Dr. K. James Hartshorn
James_Hartshorn@byu.edu

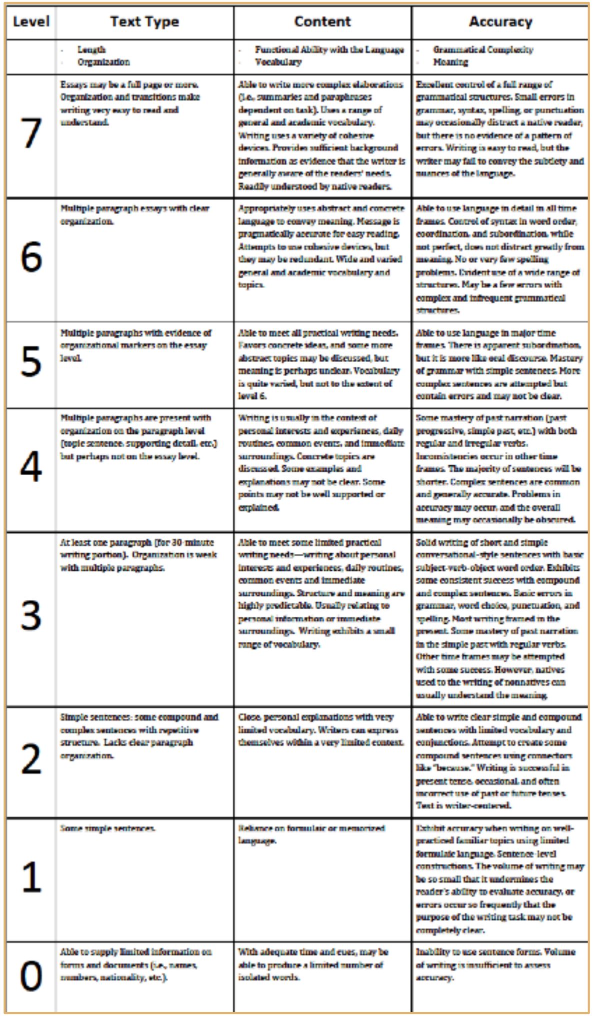
Judson Hart
hatuhart@gmail.com

essay prompt

Identify one improvement that would make your city a better place to live for people your age and explain why people your age would benefit from this change. Use specific reasons and examples to support your opinion and describe the potential immediate and long-term consequences of this improvement. You have 30 minutes to write your response.

Rubric

STUDY



Level	Text Type	Content	Accuracy
	- Length - Organization	- Functional Ability with the Language - Vocabulary	- Grammatical Complexity - Meaning
7	Essays may be a full page or more. Organization and transitions make writing very easy to read and understand.	Able to write more complex elaborations (i.e., summaries and paraphrases dependent on task). Uses a range of general and academic vocabulary. Writing uses a variety of cohesive devices. Provides sufficient background information as evidence that the writer is generally aware of the readers' needs. Readily understood by native readers.	Excellent control of a full range of grammatical structures. Small errors in grammar, syntax, spelling, or punctuation may occasionally distract a native reader, but there is no evidence of a pattern of errors. Writing is easy to read, but the writer may fail to convey the subtlety and nuances of the language.
6	Multiple paragraph essays with clear organization.	Appropriately uses abstract and concrete language to convey meaning. Message is pragmatically accurate for easy reading. Attempts to use cohesive devices, but they may be redundant. Wide and varied general and academic vocabulary and topics.	Able to use language in detail in all time frames. Control of syntax in word order, coordination, and subordination, while not perfect, does not distract greatly from meaning. No or very few spelling problems. Evident use of a wide range of structures. May be a few errors with complex and infrequent grammatical structures.
5	Multiple paragraphs with evidence of organizational markers on the essay level.	Able to meet all practical writing needs. Favors concrete ideas, and some more abstract topics may be discussed, but meaning is perhaps unclear. Vocabulary is quite varied, but not to the extent of level 6.	Able to use language in major time frames. There is apparent subordination, but it is more like oral discourse. Mastery of grammar with simple sentences. More complex sentences are attempted but contain errors and may not be clear.
4	Multiple paragraphs are present with organization on the paragraph level (topic sentence, supporting detail, etc.) but perhaps not on the essay level.	Writing is usually in the context of personal interests and experiences, daily routines, common events, and immediate surroundings. Concrete topics are discussed. Some examples and explanations may not be clear. Some points may not be well supported or explained.	Some mastery of past narration (past progressive, simple past, etc.) with both regular and irregular verbs. Inconsistencies occur in other time frames. The majority of sentences will be shorter. Complex sentences are common and generally accurate. Problems in accuracy may occur, and the overall meaning may occasionally be obscured.
3	At least one paragraph (for 30-minute writing portion). Organization is weak with multiple paragraphs.	Able to meet some limited practical writing needs—writing about personal interests and experiences, daily routines, common events and immediate surroundings. Structure and meaning are highly predictable. Usually relating to personal information or immediate surroundings. Writing exhibits a small range of vocabulary.	Solid writing of short and simple conversational-style sentences with basic subject-verb-object word order. Exhibits some consistent success with compound and complex sentences. Basic errors in grammar, word choice, punctuation, and spelling. Most writing framed in the present. Some mastery of past narration in the simple past with regular verbs. Other time frames may be attempted with some success. However, natives used to the writing of nonnatives can usually understand the meaning.
2	Simple sentences: some compound and complex sentences with repetitive structure. Lacks clear paragraph organization.	Close, personal explanations with very limited vocabulary. Writers can express themselves within a very limited context.	Able to write clear simple and compound sentences with limited vocabulary and conjunctions. Attempt to create some compound sentences using connectors like "because." Writing is successful in present tense, occasional, and often incorrect use of past or future tenses. Text is writer-centered.
1	Some simple sentences.	Reliance on formulaic or memorized language.	Exhibit accuracy when writing on well-practiced familiar topics using limited formulaic language. Sentence-level constructions. The volume of writing may be so small that it undermines the reader's ability to evaluate accuracy, or errors occur so frequently that the purpose of the writing task may not be completely clear.
0	Able to supply limited information on forms and documents (i.e., names, numbers, nationality, etc.).	With adequate time and cues, may be able to produce a limited number of isolated words.	Inability to use sentence forms. Volume of writing is insufficient to assess accuracy.