



## Doing the 'right thing' by students: Enhancing an online rater-training programme



THE UNIVERSITY  
OF AUCKLAND

NEW ZEALAND

Te Whare Wānanga o Tāmaki Makaurau

Rosemary Erlam, Janet von Randow



EALTA

Siena

5th – 8th May, 2011

# Explaining DELNA

- Diagnostic English Language Needs Assessment (DELNA)
- A post entry requirement for all first year students enrolled at the University of Auckland

- Screening

Text editing (cloze elide) &  
Vocabulary assessments

- Diagnosis

Listening comprehension






Reading comprehension

**Academic writing**

# Training raters to rate the academic English writing task

- Face-to-face training twice a year to increase inter- and intra- rater reliability
- Difficulty getting all raters together at once
- Batches of scripts come at irregular intervals through a semester so refresher training needed
- Online training program developed in 2003, trialled in 2003-2004 (Elder et al.) and changes to the program made
- Now training raters in other locations
- Can online training be used exclusively?
- What needs improving and updating?

# The Study

-  6 expert DELNA raters
-  8 novice DELNA raters
-  Rated the same 6 scripts representing the 6 DELNA Bands (4-9) online
-  Gave a simultaneous verbal protocol
-  Focus group for novice raters

	<b>Novice</b>	<b>Expert</b>
<b>Gender</b>	M1 F 7	M2 F4
<b>Age</b>	-45	46+
<b>L1</b>	English 3 Portugese 2 Japanese 1 Mandarin 1 Polish 1	English 6
<b>Other languages</b>	7 one or more other languages – almost all to a high proficiency	6 one other language 1 to a high proficiency
<b>Language Teaching and testing</b>	7	6
<b>Academic qualification</b>	PhD 4, MA 2, BA 2	MA 5 BA 1

New Zealand

The University of Auckland

## DELNA Rater Training

[Logout](#)

Welcome Janet v Randow

### Before You Begin

The following resources will prepare you for rating. Please read these before you begin and print the material you will refer to while training.

These resources are available in PDF format. You will need the free Adobe Reader (previously Acrobat Reader) to view these files - if you cannot view the files please [download Adobe Reader here](#).

- [Online Training Instructions 2005](#)
- [Rater Protocol](#)
- [Report Writing Band Descriptors](#)
- [Short Analytic Writing Marking Sheet](#)

### Task Instructions

The instruction pages of the five task versions are available for reading in PDF format:

- [Version 1](#)
- [Version 5](#)
- [Version 7](#)
- [Version 8](#)
- [Version 9](#)

### Reference Ratings

The following links show reference ratings for the DELNA bands 4 to 9.

- [Band 4](#)
- [Band 5](#)
- [Band 6](#)
- [Band 7](#)
- [Band 8](#)
- [Band 9](#)

## DELNA Report Writing Band Descriptors

FLUENCY	9	8	7	6	5	4
<b>Coherence:</b> Organisation of whole text: logical paragraphing, topic sentences, reader engagement	Very well organised; logical paragraphs, topic sentences, clear visual introduction & conclusion, message followed effortlessly	Well organised; logical paragraphs, topic sentences, relevant introductory statement, may be no concluding statement, message clear	Writing organised into paragraphs, message generally clear	Some organisation; possibly illogical/ no visual paragraphs, reader may need to reread to follow	Poor organisation; illogical/ splintered or no paragraphing, rereading does not clarify	Lacks organisation; insufficient length, and substance, incomprehensible
<b>Cohesion inside text:</b> Use of cohesive devices & referencing ('these' and 'those')	Skilful use of cohesion and referencing	Appropriate non-formulaic cohesion, some good use of referencing	Adequate or slight overuse/ formulaic use of cohesive devices, some referencing	Insufficient/ incorrect use of cohesive devices, little/ absent or repetitious referencing	Lacks cohesion; may be repetition of simple connectives (and, but, because)	No cohesion; short/ disjointed script
<b>Style:</b> Formal, academic tone hedging, appropriately impersonal, No abbreviations	Academic; 'stylish', wholly appropriate, good use of hedging devices	Academic; competent, some hedging, may be slight informality	Most aspects of academic style; no or little variation in hedging, may be some informality	Some evidence of academic style; may be overuse 1 <sup>st</sup> person/ abbreviations/ wordy/ informal language	Little understanding of academic style; language informal/ colloquial	No indication of academic style
CONTENT	9	8	7	6	5	4
<b>Part 1 of question</b> <b>Description of data so that reader can visualise graph; time place, clarity, detail</b>	Accurate, very clear description of all data; significant figures and trends	Data described clearly & accurately; includes most significant figures and all trends or vice versa	Data described accurately; includes some significant figures and trends	Data description lacks clarity; under/over emphasis on figures / misreading of some figures or trends	Data partially described; may be one piece of information only, and or inaccuracy/ confusion or repetition	Data not described or incomprehensible
<b>Part 2 of question</b> Interpretation of data - key aspects explained	Interpretation logical and wholly appropriate; relevant reasons with elaboration	Interpretation sufficient; reasons given with adequate elaboration	Interpretation generally adequate; at least 2 relevant reasons but more elaboration needed	Interpretation brief; may include relevant and irrelevant reasons	Interpretation very brief/ irrelevant/ inaccurate	Interpretation unclear or absent
<b>Part 3 of question</b> Development/ extension of ideas	Ideas (3+) relevant and supported in detail	Ideas relevant & supported	Ideas generally relevant with some supporting evidence	Ideas not always clear / supported appropriately	Lacks relevant ideas; brief/ repetitive/ unclear	No development of ideas/ incomprehensible
FORM	9	8	7	6	5	4
<b>Sentence structure</b> Variety, complexity, punctuation	Sophisticated control of sentence structure	Controlled and varied sentence structure	Satisfactory variety; occasional errors in complex sentences	Limited range; errors in complex sentences may be frequent	Limited control; mainly simple/ incomplete sentences, lacks punct.	Lacks control of sentence structure
<b>Grammatical accuracy</b>	Virtually error free	No significant errors	A few minor/repeated errors	Some significant basic errors	A range of significant errors affect expression of ideas	Short script with repeated basic errors
<b>Vocabulary &amp; spelling</b> Range and sophistication- misspelling cf. major spelling errors	Sophisticated vocabulary, wide range, appropriately used May be 1-2 misspellings	Vocab appropriate and good range May be 2-3 minor misspellings	Vocab generally appropriate, but lacks range. May be 2-3 spelling errors	Vocabulary simple, inaccurate/ inappropriate. Spelling errors intrusive	Range and use of vocabulary inadequate. Numerous errors in word formation and spelling	Basic errors in word formation/ spelling disproportionate to length and complexity of script

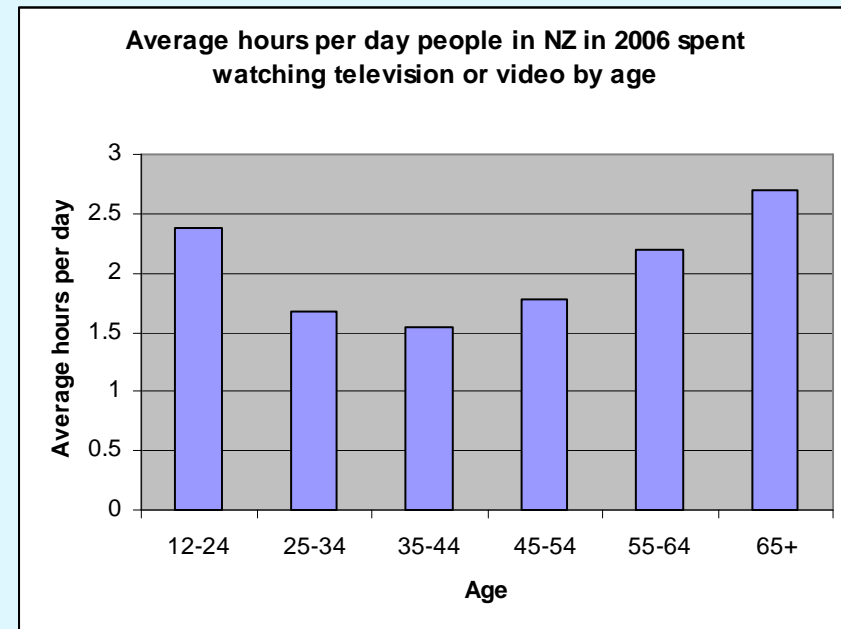
# Task

## Academic Writing

*You have 30 minutes for this task. You should write between 200 and 250 words (approx. one and a half to two pages). All sections are of equal importance.*

## Television and Video Viewing in NZ

- ▣ The graph shows the average hours per day people in New Zealand spent watching television or video in 2006.
  - ▣ Describe the information in the graph.
- THEN**
- ▣ Suggest reasons for the trends.
- AND**
- ▣ Comment on the benefits and the drawbacks that can result from watching television.





## DELNA Rater Training

### Task: Version 1

Please read the script below then complete the rating form at the bottom of this page. When you have completed the form, submit your rating to compare it with the DELNA rating of this script.

[Version 1 Instruction Page](#) (PDF)

[Back](#)

I am not properly understanding the given table. ~~A~~  
Now I am trying to explain as much as I can understand.  
~~In 1996, the average age was 27~~ Now in New Zealand, number  
of ~~states~~ child is decreasing, and the rate of population is also  
decreasing, that's mean in New Zealand now a days married people  
do not like child, it can be a good/bad things. But the very  
good and interesting things is that the average age of the people  
is increasing now, so it's mean, now medical treatment is going  
better. Now people very serious about their ~~to~~ health. And  
people also very good knowledge about family planning, They  
are very alert. In New Zealand the food ~~quality~~ quality and  
weather is very good for living and in here government is very  
serious about people's health and they are giving some useful advice  
to the people, so that people also can think about that, sometimes

## Rate the Script

Fill in the form below with your rating of this script:

### Fluency

Organisation

Comment

Band

Cohesion

Comment

Band

Style

Comment

Band

### Content

Description of data

Comment

Band

Interpretation of data

Comment

Band

Ideas

Comment

Band

**Form**

Sentence Structure

hard to find one accurate sentence. problems with simple sentence structures and boundaries, hence punctuation problems (commas). Imperative forms in lieu of discursive writing in benefits section of part 3, ditto for personal recommendation format re drawbacks.

Limited control. Short error-ridden structures. Many incomplete sentences often with repetitive structures.

Band 4

Band 5

Grammatical accuracy

Errors in every sentence - identifying subject, relative clauses/pronouns, complements (gerunds/infinitives, prepositions, articles, determiners, modals, word forms. Some correct imperative clauses

Grammar errors significant and frequent and affect the expression of ideas, e.g., tenses, articles, relative pronouns, subject repetition and prepositions.

Band 5

Band 5

Vocabulary and spelling

basic simple vocab is accurate and generally spelling is OK but ltd range (repetition of "spend time doing", "watch tv") some errors of collocation "the most of time", "to take their time" or form "informations", "so many time", "your healthy", "lock/lack"

Range and use of vocabulary inadequate and very poor punctuation.

Band 6

Band 5

***Total Band 5***

***Total Band 5***

## Summary

The summary compares your rating to the DELNA rating. Each band is compared and the difference is displayed below. A discrepancy is given for each band and a total discrepancy at the bottom of the table.

For each 1 band difference you get 1 discrepancy point, for each 2 bands difference you get 4 discrepancy points, 3 bands gets 7 points, 4 bands gets 10 points and so on. You should aim to get as low a discrepancy as possible.

	<b>Your Band</b>	<b>DELNA Band</b>	<b>Discrepancy</b>
<b>Fluency</b>	4	5	1
<b>Content</b>	5	5	0
<b>Form</b>	5	5	0
<b>TOTAL DISCREPANCY</b>			<b>1</b>

### Comment

style - my 4 your 5 - you don't really give a reason why it's a 5 over and above a 4? To me there is no indication of understanding "academic" style. Your examples are all indicative of what is wrong with this but there's no comment on what is 'right' academically.

vocab - why do you bring in punctuation as a comment here rather than under sentence structure?

# Research question 1

- ❑ Do **novice** raters, trained to rate through an online rater training program, rate as reliably as **expert** raters, trained in a face to face context?

# Pearson's correlations

More reliable		Less reliable	
Novice 2	.98	Novice 1	.84
Expert 4	.94	Expert 6	.82
Novice 8	.94	Novice 4	.82
Expert 2	.93	Novice 3	.82
Expert 1	.92	Expert 3	.78
Novice 7	.91	Novice 5	.72
Expert 5	.90	Novice 6	.70

## Diapositiva 14

---

**r1**

note that these are correlations using final bands

rerl001; 13/04/2011

	Same rating as benchmark		1 from benchmark		2 from benchmark		3 from benchmark	
	Experts	Novices	Experts	Novices	Experts	Novices	Experts	Novices
Fluency 1	58.33%	54.17%	33.33%	33.33%	5.56%	6.25%	2.78%	6.25%
Fluency 2	52.78%	35.42%	38.89%	50.00%	2.78%	6.25%	5.56%	8.33%
Fluency 3	58.33%	52.08%	41.67%	35.42%	-	12.50%	-	-
Content 1	52.78%	39.58%	41.67%	50.00%	5.56%	8.33%	-	2.10%
Content 2	50.00%	39.58%	44.44%	41.67%	2.78%	14.58%	2.78%	2.10%
Content 3	47.22%	50.00%	47.22%	39.58%	5.56%	10.42%	-	-
Form 1	63.89%	50.00%	33.33%	41.67%	2.78%	8.33%	-	-
Form 2	52.78%	52.08%	44.44%	37.50%	2.78%	10.42%	-	-
Form 3	47.22%	37.50%	44.44%	50.00%	5.56%	12.50%	2.78%	-



# Other ways of looking at the data

New Zealand

The University of Auckland

Percentage agreement		Percentage agreement	
Novice 2	100%	Novice 1	33%
Expert 4	67%	Expert 6	33%
Novice 8	67%	Novice 4	33%
Expert 2	50%	Novice 3	50%
Expert 1	17%	Expert 3	33%
Novice 7	17%	Novice 5	33%
Expert 5	50%	Novice 6	33%



New Zealand

The University of Auckland

Cohen's kappa

Novice 2

.800

Expert 4

.533

Novice 8

.528

Expert 2

.461

Expert 1

.258

Novice 7

.264

Expert 5

.336

Cohen's kappa

Novice 1

.114

Expert 6

.256

Novice 4

.185

Novice 3

.176

Expert 3

.203

Novice 5

.194

Novice 6

.007

# Research question 1

- ❑ Do **novice** raters, trained to rate through an online rater training program, rate as reliably as **expert** raters, trained in a face to face context?

# Research question 2

❑ What are the rating behaviours that distinguish **novice** raters from **expert** raters?

# Procedure . . .

## ☑ Who?

- ☑ 8 novice raters
- ☑ 6 expert raters

## ☑ Doing what?

- ☑ Rating the same 6 scripts online
- ☑ Giving a simultaneous verbal protocol

# CODING...

## ▣ Engagement with script

Citing script  
Reading script  
Rereading script

## ▣ Pre rating

## ▣ Commenting on fluency

Commenting on coherence  
Commenting on cohesion  
Commenting on style

## ▣ Commenting on content

Commenting on description  
Commenting on extension  
etc.

## ▣ Rating

Changing band  
Giving band for a particular  
category  
Justifying reason for a band  
Querying difference between bands  
Querying meaning of a band  
Referring to descriptors  
Second guessing DELNA band  
Suggesting a band  
Suggesting rating between bands

etc etc.

# nVivo analysis

- ❑ What are the rating behaviours that distinguish **novice** raters from **expert** raters?

# Some similarities to start with . . .

<b>Commenting on fluency</b>		<b>Novice</b>	<b>Expert</b>
Commenting on coherence		14.4	12.2
Commenting on cohesion		9.8	10.2
Commenting on style		11.0	11.7
<b>Commenting on content</b>			
Commenting on description		11.0	9.0
Commenting on extension		9.4	8.5
Commenting on interpretation		11.3	10.0
<b>Commenting on form</b>			
Commenting on grammatical accuracy		11.9	11.7
Commenting on sentence structure		10.6	12.3
Commenting on vocab and spelling		13.1	10.3



# Orientation to rating

New Zealand

The University of Auckland

	Novice	Expert
Expressing difficulty or inadequacy rating	5.4	1.2
Commenting on own leniency or harshness	3.8	0.3

# Engagement with script

New Zealand

The University of Auckland

	Novice	Expert
Citing script	27.6	52.8
Reading script	35.9	8.49

# Rating . . .



THE UNIVERSITY  
OF AUCKLAND

NEW ZEALAND

Te Whare Wānanga o Tāmaki Makaurau

New Zealand

The University of Auckland

	Novice	Expert
Referring to descriptors	34.0	9.3
Suggesting rating between bands	4.9	0.3

# Post rating



THE UNIVERSITY  
OF AUCKLAND

NEW ZEALAND

Te Whare Wānanga o Tāmaki Makaurau

New Zealand

The University of Auckland

	Novice	Expert
Agreeing with DELNA band	2	2
Challenging DELNA band	0.9	2.7
Commenting on problems with program	1.6	3.3

# Research question 3

❑ What are the rating behaviours that distinguish more reliable from less reliable raters?

# Two groups of raters

More reliable		Less reliable	
Novice 2	.98	Novice 1	.84
Expert 4	.94	Expert 6	.82
Novice 8	.94	Novice 4	.82
Expert 2	.93	Novice 3	.82
Expert 1	.92	Expert 3	.78
Novice 7	.91	Novice 5	.72
Expert 5	.90	Novice 6	.70

## Diapositiva 29

---

r2

note that these are correlations using final bands

rerl001; 13/04/2011

# Orientation

New Zealand

The University of Auckland

	More reliable	Less reliable
Apologising	0	1.3
Expressing difficulty or inadequacy rating	1	6.1



# Engagement with script



THE UNIVERSITY  
OF AUCKLAND

NEW ZEALAND

Te Whare Wānanga o Tāmaki Makaurau

New Zealand

The University of Auckland

	More reliable	Less reliable
Citing script	55.7	21.1
Reading script	28.9	19.4
Comparing scripts	1.7	0.1

# Post rating



THE UNIVERSITY  
OF AUCKLAND

NEW ZEALAND

Te Whare Wānanga o Tāmaki Makaurau

New Zealand

The University of Auckland

	More reliable	Less reliable
Agreeing with DELNA band	1	3
Challenging DELNA band	2.9	0.4
Commenting on discrepancy between self & benchmark	3.4	7
Commenting on problems with program	2.7	1.1

# Technical changes needed post 2005

- More detailed comments on the DELNA side of the rating grid
- A glossary of terms especially for raters in other locations
- A split screen with script on one side and the rating grid on the other
- Updating the rating template by adding a 3rd column where raters can keep a tally of form errors
- A chance to go back and relook at the script post submitting and comparing ratings
- Compulsory entering of comment after comparing ratings

# References (1)

- ❑ Charmaz, K. (2006) *Constructing Grounded Theory. A Practical Guide Through Qualitative Analysis*. London, Sage Publications Ltd.
- ❑ Elder, C., Knoch, U., Barkhuizen, G. and von Randow, J. (2005) Feedback to enhance rater training. Does it work? *Language Assessment Quarterly* 2, 3, 175-196.
- ❑ Elder, C., Barkhuizen, G., Knoch, U. and von Randow, J. (2007) Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing* 2007; 24; 37
- ❑ Fox, Janna D. (2003). From products to process: An ecological approach to bias detection. *International Journal of Testing*, 3 (1).
- ❑ Gass, S.M. & Mackey, A. (2000) *Stimulated Recall Methodology in Second language Research*. Mahwah, NJ 07430, Lawrence Erlbaum Associates, Inc.
- ❑ Knoch, U., Read, J. and von Randow, J. (2007) Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing* 12, 26–43
- ❑ Knoch, U. (2009) *Diagnostic Writing Assessment: The Development and Validation of a Rating Scale*. Frankfurt am Main, Peter Lang.

# References (2)

- ❑ Lazaraton, A., and Taylor, L. (2007) *Qualitative Research Methods in Language Test Development*. In Fox, J., Wesche, M, Bayliss, D., Cheng, L., Turner, Carolyn, E. And Doe ,C. *Language Testing Reconsidered* (pp113-130) Ottawa, University of Ottawa Press.
- ❑ Lumley, T.J.N. *The process of assessment of writing performance: the rater's perspective*. Unpublished PhD, Department of Linguistics and Applied Linguistics, The University of Melbourne.
- ❑ McNamara, T.F. (2000) *Language Testing*. Oxford, Oxford University Press.
- ❑ Teddlie, C. & Tashakkori, A. (2009) *Foundations of Mixed Methods Research. Integrating Quantitative and Qualitative Approaches in the Social and Behavioural Sciences*. Thousand Oaks, California. Sage Publications Inc.