Investigating an online rater training program: product and process

Rosemary Erlam
Janet von Randow
John Read
The University of Auckland

The Diagnostic English Language Needs Assessment (DELNA) programme at the University of Auckland has traditionally offered face-to-face training to both new raters and to experienced raters of writing who need to undergo refresher training in order to increase inter-rater and intra-rater reliability. However, for reasons of practicality and convenience, an online program to offer refresher training for experienced raters was developed and revised after trialling in 2003-04 (Elder, Barkhuizen, Knoch & von Randow, 2007).

The study reported on in this paper arose out of the need to establish the validity of using this online DELNA program to train new raters (i.e., novice raters). Rater-training outcomes were investigated through the analysis of quantitative data. However, there was also a focus on the 'process of rating' using qualitative data, that is, think-aloud protocol, to identify rating behaviours associated with more vs. less reliable ratings.

Results suggest that this program has potential to train novice raters to rate reliably but that future research needs to investigate more rigorously the impact of a program like this one on subsequent rating. Nvivo analysis of the think-aloud protocols highlighted key aspects of the rating behaviours of more and less reliable raters.

Key words: rater training, rater reliability, writing assessment, online, rater behaviour

Address for correspondence: Rosemary Erlam, Department of Applied Language Studies and Linguistics, The University of Auckland, Private Bag 92019, Auckland 1010, New Zealand; Email: r.erlam@auckland.ac.nz.

Background

DELNA is a diagnostic assessment, first developed in 2001 (Elder & Erlam, 2001), which aims to identify students with academic English language proficiency needs and to direct them to language support as appropriate. All first year undergraduate students and 'new' PhD students admitted to the University of Auckland are now required to complete a DELNA assessment. DELNA consists of two phases. The first of these, which all students complete online, is known as the DELNA 'Screening' and consists of speeded reading and vocabulary tasks. The second is completed by those students who fall below the cut-score on the Screening. It is known as the DELNA 'Diagnosis' and includes paper-based reading, listening and writing tasks (for a more complete description, see Read, 2008). This paper focuses on the rating of the writing assessment task. While DELNA was designed as a low-stakes assessment tool, it is important for the credibility of the programme, and the quality of the advice given to the students that the rating of the writing task should be as reliable as possible. Traditionally DELNA raters have been trained to rate scripts in a face-to-face training session where they are given a series of benchmarked scripts to rate independently, following by discussion and feedback from their peers and from a DELNA facilitator. This has been followed by independent rating of a larger batch of scripts and monitoring of score reliability.

The following literature review will briefly discuss the importance of rater training and refresher rater training in reducing variability in rating behaviour. It will then present some early initiatives in self-training and in later online rater training. Two online rater training studies conducted in the DELNA context which foreground the present study are discussed in some detail. The second main section of the literature review differentiates a product and process approach to evaluation of rater training and presents the advantages and limitations of think-aloud protocols, the qualitative data that informs the present study. Finally there is reference to the literature that investigates variation in rater behaviour that occurs as a result of differences in rating experience and in the ability to rate proficiently or reliably.

Rater Training

The subjective nature of evaluations of writing quality by raters is a potential threat to test fairness (Elder et al., 2007). An important way of ensuring the quality of rater-mediated assessment is to provide both initial and on-going training for raters. A number of studies have provided evidence to suggest that rater-training can be effective in reducing variability in rating behaviour, in

improving rater reliability and in reducing rater bias (e.g., McIntyre, 1993; Weigle, 1994, 1998). However, there is also evidence to suggest that the effects of rating may last for a limited period of time (Congdon & McQueen, 2000), a finding which highlights a need for ongoing 'refresher' rater training.

The need for the training of raters in diverse locations as well as the need for the regular 'retraining' of raters has led to an interest in the possibility of online rater training. A self-training kit for raters of the SOPI (Simulated Oral Proficiency Interview) was an early response to the issue of geographical dispersion (Kenyon & Stansfield, 1993) as were the self-instruction manuals made available for both the writing and speaking sections of IELTS for examiners in distant locations. The first online program, however, was developed for the training of raters of writing in the English Language Centre at the Hong Kong Polytechnic University (Hamilton, Reddel & Spratt, 2001). An investigation into raters' responses to the program demonstrated that it was theoretically supported but that its application was less successful. Hamilton et al. concluded that an online training initiative would need to include the opportunity for raters to discuss, for example, in a discussion forum, their assessment decisions with other raters.

Two studies have investigated the impact of online training for raters of writing within the context of DELNA, and thus have more directly informed the study reported on in this paper. In both these studies an analytic rating scale was used, unlike other research in online training which has had raters work with holistic scales (Hamilton et al., 2001; Wolfe, Kao & Ranney, 2010). There is some evidence to suggest (Francis, 1977 & Adams 1981, both cited in Weir, 1990) that analytic scoring is more useful for training inexperienced raters as they can more easily understand and apply the criteria than in a holistic scale. The first of the DELNA studies (Elder et al., 2007) looked at whether using an online program can train new DELNA raters and offer refresher training for experienced DELNA raters. In this study, eight participants (six of whom were already trained and experienced DELNA raters) first rated 100 DELNA scripts in their own time. They then completed an online rater-training program where they rated at least 10 scripts online and received immediate feedback about their performance. Lastly, participants re-rated 50 of the original 100 scripts, again in their own time. FACETS analyses of participant ratings for the 50 scripts that were rated both before and after training, showed slightly higher levels of overall inter-rater agreement after online training and, in some instances, reduced levels of inconsistency and bias.

In a subsequent study, Knoch, Read and von Randow (2007) used the modified version of this same program. Changes included a discrepancy score giving

raters information about how their rating compared to the benchmark in each analytic rating category and a series of Reference Ratings with detailed comments that raters would view before beginning rating. Knoch et al. (2007) then compared the effectiveness of face-to-face and online rater-training in the re-training of experienced DELNA raters. First, the 16 participants in the study rated the same 70 scripts and then they were divided into two groups of eight. One group rated a further 15 scripts online and received online information about discrepancies between their ratings and the benchmark ones, while the other group rated the same scripts in a face-to-face context and received individualised, face-to-face feedback about their rating based on the results of a FACETS analysis. Both groups then rerated the initial 70 scripts. The author concluded from the results that, overall, both types of training were effective but that maybe the online training was slightly more successful in encouraging raters to be more consistent in their rating behaviour. The face-to-face training appeared marginally more effective in reducing individual biases, perhaps because raters in this condition received individualised feedback whereas those in the online rating condition did not.

The results of Knoch et al (2007) contrast, however, with the findings of a larger scale study conducted by Brown and Jaquith (2007), also comparing the efficacy of online and face to face rater training. A Many-facet Rasch analysis found that most of the extreme (in terms of severity and leniency) and unexpected responses came from the online trained raters. Brown and Jacquith (2007) call for further research on online rater training, in particular to investigate the impact of this mode of marking on rating quality.

The present study aims to further investigate the suitability of training raters of writing online, in building, in particular, on the previous two studies conducted in a DELNA context. It investigates whether this DELNA online training program can now be used to train novice (inexperienced) raters in addition to its role in refreshing rater-training for expert raters.

Product and process

An evaluation of a rater-training program will usually look at whether the program outcomes are effective in terms of 'product' (Fox, 2003, p. 21). This usually involves the comparison of rater scoring data to establish rater reliability (e.g. before and after training). However an evaluation of rater-training also needs to consider to what extent scoring procedures are being implemented in an appropriate way, an investigation that will involve a study of rater behaviour through the analysis of qualitative data (Weigle, 2002). This involves an evaluation of the rating 'process'.

A mixed-methods research design allows for the collection of both quantitative and qualitative data in order to achieve a more complete understanding and extend the breadth of the investigation (Dörnyei, 2007). Quantitative data usually comprises information about the reliability of scores/grades or the consistency/severity of rater behaviour. On the other hand, qualitative data typically comprises verbal report protocols to allow an in-depth investigation of the rating process. O'Hagan (2010) points out that there have been a number of studies investigating rater cognition using verbal reports in both L1 and L2 writing assessment (e.g. Cumming, 1990; Weigle, 1994), which have contributed both to an understanding of influences on rater behaviour and have helped shape best practice in verbal report research.

O'Hagan (2010) argues that two important advantages of the verbal report (also referred to as think-aloud protocol), are firstly, its directness, that is, it gives information about actual instances of behaviour, and, secondly, the immediacy of the link between data elicitation and the object of investigation. Wigglesworth (2005) also maintains that verbal reports allow us probably the most direct insight into thought processes. However, there have been concerns that the constraints of giving a protocol may disrupt and thus falsify the account of the rating process (Lumley, 2005; Stratman & Hamp-Lyons, 1994). Barkaoui (2011) maintains that it is important, in using think-aloud, to realise its possible incompleteness as a data-gathering tool. He raises the issue of veridicality, that is, the extent to which the comments provided during the think-aloud are an accurate representation of the raters' cognitive processes during rating. He cautions that lack of mention of a particular feature does not mean that it was not in evidence, nor does mention of a feature equate with importance. He also discusses reactivity, defined by Loewen and Reinders (2011) as the extent to which performing a task for research purposes alters the nature of the task. Barkaoui claims that it is important to realise that think-aloud may alter the rating process, providing evidence in his own research to show that more experienced raters had more difficulty rating while thinking aloud than did less experienced raters. He concludes, however, that the use of think-aloud protocols can still be defended, provided that they are used with caution. They do give insight into the kinds of processes used in rating and, perhaps more importantly, essay rating may be less susceptible to the negative effects of think-aloud because raters normally have to justify their scores and are, also, very often aware of an audience when rating.

The investigation of rater behaviour during rating has been a fertile area of research (e.g. Lumley, 2005; O'Hagan, 2010; Schoonen, 2005; Weigle, 1998). Eckes (2008, p. 181) maintains that better understanding and explaining rater variability is 'one of the biggest challenges for language assessment researchers

to date'. A key area of research interest to date is the investigation of the impact of rating experience on rating (Pula & Huot, 1993; Schoonen, Vergeer & Eiting, 1997; Vaughan, 1991), a question that has been investigated in a number of studies (e.g. Barkaoui, 2010; Weigle,1999).

Weigle found that novice raters rated essays written in response to one of two given prompts more severely than the experienced raters but differences between the two groups were eliminated as a result of rater training. Barkaoui (2010) had experienced and novice raters rate a set of essays holistically. He found, in contrast to Weigle, that experienced raters gave lower scores and exhibited less variety in scoring than the novice raters. They also paid more attention to linguistic accuracy whereas the novices gave more weight to argumentation. Barkaoui (2010) explains that these differences could be due to significant differences in the backgrounds of the two groups.

Wolfe et al. (1998) looked at variation in rater behaviour in terms of proficiency rather than experience. Proficiency was defined in terms of the ability to maintain a high level of reliability in rating. Wolfe et al. found that raters who were more proficient at rating focused more on general features of the essays they assessed, whereas less proficient raters focused more on specific features of the essay.

In conclusion, there is some research evidence to suggest that online rater-training may be effective in reducing variability in the rating of written assessments. However, there is a lack of research that investigates this option as a possibility for training inexperienced (novice) raters ab initio. An evaluation of the effectiveness of rater training has typically involved investigation of product, or rating outcomes. Some research has also investigated the rating process in an attempt to understand the types of rater behaviours that might distinguish, for example, experienced from novice raters. There has been less of a focus in this research at looking at the types of rater behaviour that may differentiate more reliable (i.e. more proficient) from less reliable raters.

The present study

This study aims to investigate two groups of raters as they completed an online rater training program. One group (n=6) consisted of experienced DELNA raters, whereas the other group (n=8) were novices with respect to DELNA. Both groups produced think-aloud protocols as they rated a series of six scripts. The ratings allowed for investigation of the *product* of assessment while the think-aloud protocols allowed for an investigation of the *process* of assessment. The research questions were as follows:

1. Do novice raters, trained to rate through an online rater training program, rate as reliably as expert raters originally trained in a face-to-face context?
2. What are the rating behaviours that distinguish more reliable from less reliable raters?

Method

The assessment instrument

The DELNA writing task used in this study is designed to represent the type of academic writing (i.e., expository essay writing) that is often required of students at University level. The writing task involves the description and interpretation of data presented in either a graph or a table. The DELNA rating scale is used to score all writing scripts analytically according to three separate categories: fluency, content and form. Within each of these three categories, there are three subcategories (e.g. fluency comprises coherence, text cohesion and style; see Table 3 for all categories and subcategories), meaning that for each script there are nine rating decisions to be made. The scale was developed from other existing rating scales to define the underlying construct (i.e. the ability to write a clearly structured and coherent text that describes and elaborates on the given data) and has further evolved over time in response to feedback from raters and assessment specialists involved in the design and implementation of DELNA. More recently it has been further adapted in response to research conducted by Knoch (2007) in her analysis of DELNA writing performances. A scale with six band levels (from levels 4 to 9) gives descriptors for each subcategory at each level. An overall score at band level 4 signifies that the student is at severe risk of not succeeding academically, a score at level 9 implies that the student has language skills that should equip him/her well for University study. Following rating, scores within individual categories are totalled by the computer to give a score which is then divided by three to give either an exact band, or, if the total cannot be divided to give a whole number, the next lowest band.[i] The three category scores are averaged to give an overall score.[ii] All writing tasks are assessed by two raters and a third rater is used when the two raters differ by more than one DELNA band on the overall score. Students who score 6 or below for writing are given an appointment with the DELNA advisor who will discuss with them their scores in the individual categories with the aim of identifying their specific weaknesses, giving them information at the same time about where to get language enrichment support.

The raters

The expert raters (n = 6) in this study had been trained to rate DELNA writing scripts in a face-to-face context. They had all been rating DELNA scripts for between 2 to 10 years and each had rated between 250 and 2500 scripts during that time. Three of these experts were also teaching university academic writing courses and thus had ongoing experience in assessing writing. The other three experts were ESOL teachers each with over 25 years' teaching experience. At the time of the study the expert raters had not rated for DELNA for some months and were therefore due for refresher training. Novice raters (n = 8) had had no prior experience in rating for DELNA but all mostly had some background in language testing, either through formal study at a tertiary institution or from teaching experience. Six of the eight said that they had experience in rating writing and six (not the same six in every case) said they had experience teaching writing. Table 1 shows that there were some other differences between the novice and expert raters: the novice raters were considerably younger than the experts, were for the most part more qualified academically and came from a wider variety of language backgrounds than the expert raters, who were all native speakers of English.

Table 1. Rater background information

	Novice raters: n = 8	**Expert raters: n = 6**
Gender	M1 F 7	M2 F4
Age	20 – 45 years	45 – 68 years
Language identified as first language	English 3 Portuguese 2 Japanese 1 Mandarin 1 Polish 1	English 6
Other languages	7 spoke one or more other languages, in almost every case to a high proficiency	6 spoke one other language, 1 to a high proficiency
Language teaching experience	7 had 10 or less years of language teaching experience	All had language teaching experience, 5 had 15 years or more
Academic qualification	PhD 4, MA 2, BA 2	MA 5, BA 1

Procedures

The online training programme

The training involved the rating of a set of 6 benchmark writing scripts gathered from previous test administrations and representing the full range of proficiency levels (DELNA bands 4 to 9), although raters were not aware that there was a script representing each band level. These writing samples were

chosen because they had been given exactly the same bands in each category by two experienced DELNA raters and a panel of an additional 4 DELNA raters had rated and discussed them before they were put online, generating written comments that were used as benchmark comments in the training programme. The scripts were all written in response to the same prompt (Television and Video Viewing in New Zealand, see Appendix A). The training programme also included the DELNA band descriptors (hard and soft copies) and another 6 scanned scripts (written in response to a different prompt to those scripts used for rating in this study) which represented different band levels and with which raters could familiarise themselves before beginning rating.

Rater training/procedures

The novice raters were given, prior to beginning rater training, six articles to read on the subject of rating writing. Two (Elder, Knoch, Barkhuizen, & von Randow, 2005; Elder et al., 2007) of these articles described research investigating the rating process within the context of DELNA. The aim was to familiarise raters with the rating process and with the history of DELNA. They were first asked to write comments on the articles they had read in response to guided prompts. This had the aim of ensuring that the raters did read and attend to the content of the articles. Six out of the eight raters wrote comments, which were read to ensure that there were no inappropriate misperceptions with regard to the rating process. (The expert raters were not required to read any articles, given that all of them were familiar with DELNA and with rating writing in this context). On completion of the readings they were given a set of written instructions outlining the training process. This included an individual logon, password and the URL for the rater training program. They were asked to familiarise themselves with the DELNA band descriptors (they had a hard and soft copy). Instructions directed them to where on the website they could access the 6 scanned sample scripts (mentioned above) which would familiarise them with the different band levels.

The expert raters did not follow these two steps (read articles or access 6 sample scripts). Their previous training had taken place in a face-to-face training session where they were first given a series of benchmarked scripts to rate independently. They then received feedback from their peers and from a DELNA facilitator about the appropriacy of their ratings. This was followed by independent rating of a larger number of scripts and monitoring of score reliability.

As they rated each script in terms of the three categories, raters were required to write brief comments, giving reasons for each rating decision. A prompt would

remind them to do this if they tried to move on to the next category without having done so. On submitting their series of ratings for each script, raters were shown their band scores and comments in relation to the DELNA benchmark band scores and comments. This information was followed by a 'discrepancy score' indicating the discrepancy (a plus or minus indicated the direction) between their own and the benchmark band scores (see Appendix B, Screenshot 1). Raters were then encouraged to enter a written comment into a box indicating whether or not they agreed with the benchmark and making any other comment they wanted to about the rating process. The raters had no information about the writers of the six scripts, other than that they were all first year undergraduate University students.

Think-aloud training

Before beginning rating all raters, both novice and expert, were given practice in recording a think-aloud protocol whilst rating (the scripts chosen for this purpose were written in response to a different prompt from that used in this study). They were told that they were to record their thoughts as they rated – 'try to speak your thoughts aloud into the microphone WHILE you are rating the script'. When they had rated a couple of scripts, the researcher checked them and, once she had established that they were, indeed, fulfilling the requirement of think-aloud protocol correctly, she authorised them to begin rating the six benchmarked scripts selected for this project and at the same time to record think-aloud protocols. The raters were advised that this could well take more than one session (i.e., approximately two hours) and indeed, most raters took two sessions over two consecutive days to finish rating all six scripts.

Summary of procedures for novice and expert raters

To sum up, the training experience differed for the novice and expert raters in the following ways:

1. The novice raters were asked to read a series of research papers, whereas the experienced raters were not.
2. The novice raters were required to spend some time familiarising themselves with the DELNA writing descriptors and with six sample scripts before beginning to rate the six benchmarked scripts chosen for inclusion in this study; this was not considered necessary for the experienced raters, who had already had considerable experience rating for DELNA.

Following these initial stages, the procedure was the same for both groups in terms of recording the think-aloud protocol, rating the six scripts and receiving feedback on rating. It is important to reiterate that this process was considered 'refresher training' for the expert raters and initial training for the novice raters. One important feature that differentiated this online training from the face-to-face training that the expert raters had received as novice raters was the immediacy of the provision of feedback. The online program gave raters individual information about how their ratings compared with benchmark scripts immediately following the rating of each script. This feature had been a component of the previous research investigating online rater training (Elder et al, 2007; Knoch et al, 2007). In contrast, the face-to-face training context had not provided raters with feedback until they had rated all scripts and then this information was shared in the context of a discussion with their peers.

Collection of data

This study adopted a mixed methods design. The primary data set comprised quantitative data, that is, the raters' ratings as they each assessed the six scripts that had been chosen for this study and qualitative data in the form of the think-aloud protocols. The qualitative data also included written comments that some participants made as they read the DELNA benchmarks after rating each script. Some raters chose to respond in written form to the benchmarks in the box provided for that purpose in the online template, whereas others were happy to voice their reactions at the end of the think aloud protocols. It was decided that the data set would be more complete in this instance if both types of data were taken into account.

Analysis of quantitative data

The quantitative data were entered into SPSS (Version 18) along with the DELNA benchmark ratings to allow a number of comparisons to be made. In the first of these, Spearman's correlations were conducted between raters' scores and the DELNA benchmark ratings. A secondary analysis calculated percentage agreement of raters' individual band ratings in each category with the DELNA benchmark ratings. Finally Cohen's weighted kappa established agreement of final bands for each script with DELNA final bands.

Analysis of qualitative data

The think-aloud protocols were all transcribed by the same research assistant who had been seconded to work on the project. Two of the researchers then chose one script and, using a grounded theory approach, analysed the 14 think-aloud protocols that the raters had generated as they rated this script. The approach to developing coding categories was data-driven (O'Hagan, 2010) in

that researchers looked for the salient themes or patterns emerging from the data (Ellis & Barkhuizen, 2005). They initially developed a set of coding categories independently as they read the 14 protocols and identified what appeared to be the common themes, constantly refining them as they worked with them in what O'Hagan (2010) describes as a recursive process. The researchers next shared their coding categories and selected those that they had in common and those which seemed to best encapsulate the salient themes of the data. The final list of the 38 coding categories can be found in Appendix C. Nine of the categories corresponded directly to the DELNA rating subcategories as the researchers wanted to investigate rater attention to these during the rating process. Other categories were developed from the initial rating (described above) to investigate different stages in the rating process, for example, the *while rating* phase of 'suggesting' or 'giving a band', 'changing a band', 'justifying a reason for a band' and so on. Another focus of the analysis was each rater's affective response to the online training program, so a number of codes (e.g. 'commenting on problems with the program', 'expressing problems or difficulty rating') were motivated by this aim. The coding categories are grouped broadly according to the stages in the process, although it is important to point out that, while this grouping was useful in managing the data, it was not definitive and some coding categories were in evidence at a variety of stages in the rating process. For example, 'commenting on leniency or harshness' occurred during 'rating' or 'post rating' as well as during the 'pre rating' phase. Overall, however, the coding categories were significantly influenced by the DELNA rating process as prescribed by the descriptors. They are therefore, for the most part, different to the 28 decision making behaviours identified by Cumming (1990) (some categories in common include 'assess coherence', 'establish appropriateness of lexis' etc.). This is because, in contrast to procedure in this study, Cumming provided raters with no rating descriptors, telling them only that they were to rate scripts on a scale of 1 to 4 for 3 linguistic criteria.

Once the set of coding categories had been finalised, all the verbal protocols (n = 84) and the written comments in response to DELNA benchmarks were coded by one of the researchers using NVivo 9 (http://www.qsrinternational.com). NVivotools then allowed for analysis of the data in order to answer the research questions. For each coding category, totals representing the number of occurrences in each script were calculated for each rater group and then divided by the number of raters respectively to give averages of occurrence of each category for the two groups. A summary of this process is given in Figure 1 below.

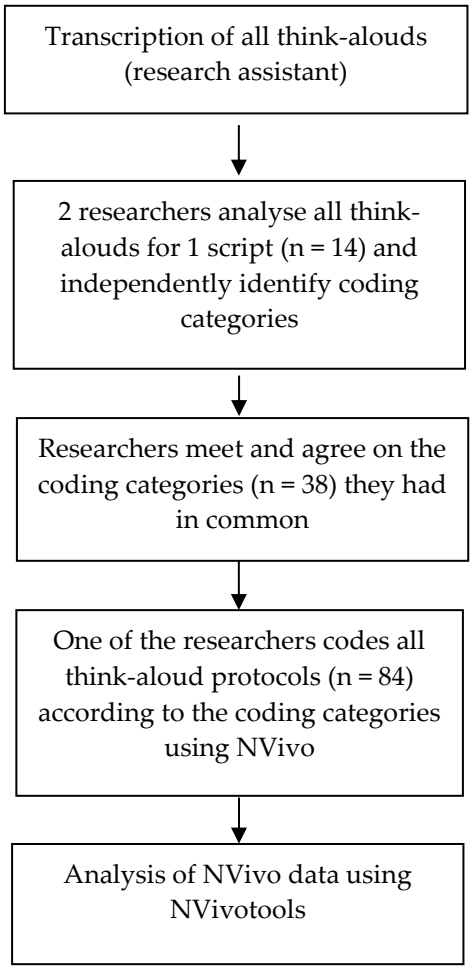


Figure 1. Summary of qualitative data analysis process.

Results and discussion

The first research question asked whether the novice raters could be trained, through this online rater training program, to rate as reliably as the expert raters.

An initial analysis involved correlating, for each of the 6 scripts, raters' total bands for each category (fluency, content, form) with the DELNA benchmark bands for each category. Results are presented in Table 2, ranked in terms of strength of correlation with the DELNA benchmark bands. They are thus presented in two separate groups – Group 1 consists of those raters who achieved a correlation of .9 or over with the benchmark ratings (Novice 7 is included, as their correlation, if rounded to one decimal point, would fit this category) and Group 2 is made up of those raters who achieved correlations of .8 or less.

Table 2. Spearman's correlations between raters' bands and DELNA benchmark bands.

Group 1		Group 2	
Novice 2	.98	Novice 3	.82
Expert 4	.95	Expert 3	.80
Novice 8	.94	Novice 1	.79
Expert 1	.93	Novice 4	.77
Expert 2	.92	Expert 6	.75
Expert 5	.90	Novice 6	.72
Novice 7	.89	Novice 5	.60

All correlations are statistically significant to 0.01.

From the results in Table 2, we can see that there is evidence to suggest that novices can be trained in this way to rate as reliably as experts; indeed, it was a novice who had the highest correlation of all (r=.98) and two of the three raters with the highest correlations were novice raters. These results paint a fairly optimistic overall view of the reliability of the online rater training program, given that half the raters had correlations of .9 or above.

It was decided that a secondary analysis should be conducted, taking raters' individual band ratings in each category, calculating percentage agreement with the DELNA benchmark ratings for each category and presenting these for each group. The percentage of ratings which differed from the benchmark ratings (in terms of discrepancies of 1, 2 or 3 bands) was also calculated. The results are presented in Table 3.

Table 3. Percentage agreement of ratings for individual categories with DELNA benchmarks

DELNA rating categories		Same rating as benchmark		1 from benchmark		2 from benchmark		3 from benchmark	
		Exp.	Nov.	Exp.	Nov.	Exp.	Nov.	Exp.	Nov.
fluency	coherence	58%*	54%	33%	33%	6%	6%	3%	6%
	cohesion	53%	35%	39%	50%	3%	6%	6%	8%
	Style	58%	52%	42%	35%	-	13%	-	-
content	description	53%	40%	42%	50%	6%	8%	-	2%
	interpretation	50%	40%	44%	42%	3%	15%	3%	2%
	extension	47%	50%	47%	40%	6%	10%	-	-
form	sentence structure	64%	50%	33%	42%	3%	8%	-	-
	gramm. accuracy	53%	52%	44%	38%	3%	10%	-	-
	vocab. & spelling	47%	38%	44%	50%	6%	13%	3%	-

*Percentages were rounded up or down to give whole numbers. Where there are no percentages, there were no ratings that corresponded to the category.

Table 3 demonstrates that there were some substantial differences between the two groups for some categories. For example, there were differences of 10% or more between percentage agreement ratings of the two groups for the ratings of cohesion, description, interpretation and for sentence structure. However, what is perhaps most notable about these data are the overall low percentage

agreement ratings of both groups with the DELNA benchmarks. Neither group scored a percentage agreement rating for any category of more than 64% with the benchmarks. Furthermore, for at least three subcategories (cohesion, description, vocab & spelling) 50% of the novice ratings differed from the DELNA benchmark ratings by one whole band score. It is interesting to note that these subcategories were evenly spread among the three overall categories of Fluency, Content and Form, that is, that the differences did not imply a particular difficulty with one rating category. For at least six categories 40% or more of the expert ratings differed from the DELNA benchmark ratings by one band score. Here it should be noted that 40% of the expert ratings for all subcategories of the Content category differed from the DELNA benchmark scale by at least one band, and that 40% of ratings for two of the Form categories also differed from the DELNA benchmark by at least one band.

A final comparison of the ratings in relation to the DELNA benchmarks was conducted taking the total band rating for each script. Cohen's weighted kappa was calculated to establish agreement of final bands (i.e. one overall band) for each script with DELNA final bands. This calculation, which corrects for chance agreement, tells us to what extent the outcome for each student would vary or be consistent with the DELNA overall rating. Weighted kappa (Cohen, 1968) weights discrepancies differently, giving more weight to greater discrepancies. For example, weighted kappa treats a discrepancy of more than 1 band from the benchmark more severely than a discrepancy of only 1 band. This reflects the fact that the greater the disagreement between the rating given and the benchmark the greater the difference in outcome for the student in terms of advice given or support offered. In this study, as in Xi and Mollaun (2011), quadratic weights are used so that 0 is assigned to perfect agreement, 1 to a discrepancy of 1 band, 4 to a discrepancy of 2 bands and 9 to a discrepancy of 3 bands. The results are presented in Table 4 with raters ranked in terms of degree of concordance.

Table 4. Overall level of agreement with final DELNA bands by rater (Cohen's weighted kappa).

Novice 2	1.0	Expert 1	.84
Novice 8	.95	Novice 4	.84
Expert 4	.94	Novice 1	.80
Expert 5	.90	Novice 3	.76
Expert 2	.89	Expert 3	.76
Expert 6	.86	Novice 5	.66
Novice 7	.84	Novice 6	.66

Results in Table 4 demonstrate that twelve raters had kappas that could be considered excellent ($K > 0.75$), while two raters (Novice 5 & 6) had kappas that

were good (0.60 < K < 0.75), using the assessment of significance Robson (2002) suggests[iii]. The high levels of concordance demonstrated by these results suggest that this rater training program has potential as an instrument to improve rater reliability, when overall bands are taken into consideration, in that all expert raters achieved kappas that were excellent. Results also suggest that this program may have possibilities as a tool to train new or inexperienced raters, in that all but two of the eight novices also had excellent kappas. With respect to Research question 1, therefore, there is evidence to show that novice raters performed similarly to the expert raters as a result of the online training program, given that two out of the three raters with the highest Kappas were novices. It is perhaps important to point out that the potential impact of background variables (e.g., different L1 backgrounds, higher level academic qualifications) that differentiate the two groups is unknown. In interpreting these results a note of caution is warranted. Firstly, it is important to remember that the results presented are based on a small number of scripts representing a range of performance where agreement is likely to be higher than it may be for a batch of scripts more in the middle of the score range. Secondly, the high levels of concordance with DELNA total band scores indicate that raters may appropriately identify students at risk but do not necessarily mean that raters may perform so well in diagnosing test taker performance in relation to the subcategories (as data in Table 3 suggests). An examination of the final band scores given for each category (fluency, content, form) in relation to benchmark bands for these categories, shows that for four of the raters with excellent kappas, at least 1 of the 6 scripts rated differed by 2 or more bands from the benchmark for two or more of these three categories. A discrepancy of this nature might result in test takers receiving very different advice about the weaknesses they have in writing when they meet one on one with the DELNA advisor. In actual practice, however, the DELNA advisors (trained DELNA raters themselves) usually look at the student's script before an advisory session and make their own judgements about the strengths and weaknesses in the student's writing. In this respect they are not relying solely on raters' judgements and so a lack of reliability in category rating may not be so crucial. Another limitation of the present study is that the research design did not include a measure of the impact of the online rater training session on subsequent rating, unlike in previous research (Knoch et al, 2007) where raters were asked to rate 70 scripts after completing online training.

Research question 2 asked what rating behaviours distinguished more reliable from less reliable raters. In order to answer this question, the raters were divided into two equal groups, one that was considered more reliable and one that was considered less reliable. The more reliable raters (i.e., Group 1 in Table

2) all had Pearson correlations of over .9 when measured against the DELNA benchmark ratings for each category (fluency, content, form). All of these seven more reliable raters also had Cohen's Kappas that could be classified excellent when total band ratings were compared with DELNA benchmark ratings. The data used to answer Research question 2 was the think aloud protocols and written comments of all raters as they rated the same six scripts. After coding, an NVivo query was conducted whereby the data was analysed according to the more/less reliable grouping. The researchers then together examined the results, presented in Appendix C, and identified those differences that appeared to be substantial.

Results for rater attention to the different subcategories of the DELNA descriptors during the rating process will be discussed first, as interest in this issue was a primary one (see Collection and analysis of data). The results are presented in Table 5. Each of the figures in the table represents how many times, on average, each rater in the respective groups made reference to each category in rating the six scripts. The more reliable raters, perhaps unsurprisingly, all made more references to the nine DELNA subcategories during the 'think-aloud' than did the less reliable raters, although on at least two occasions the difference was negligible (commenting on style, 11.9/11.1; commenting on vocab & spelling, 12.9/12.2). For some categories the difference appeared more substantial, notably for 'commenting on cohesion' (13/7.2) and, perhaps also, for 'commenting on interpretation' (13.8/9.7). Results for another subcategory of Fluency, 'commenting on coherence' (15.3/12) also showed some difference between the two groups. It was decided to investigate whether differences in the number of references was in any way associated with differences in word length for the two groups. Word length was calculated with respect to think-aloud protocols as a whole, rather than with respect to individual categories. The think-aloud protocols of the more reliable raters were on average longer (5289 words) than those of the less reliable raters (4122 words), suggesting a greater 'time on task'. Baume, York and Coffey (2004) found that greater 'time on task' in terms of time spent assessing writing led to an increase in severity of marks, quoting one rater as saying that taking longer led to more flaws being found. It is interesting to speculate in this study to what extent greater reliability may be associated with more 'time on task', affording greater opportunity to attend more closely to student writing. However, it is possible that raters in this study spent more time on task than they might in an operational, or non-research, setting.

Table 5. DELNA descriptor subcategories: mean frequency of comments by more and less reliable rater groups

Coding category	More reliable	Less reliable
commenting on fluency		
commenting on coherence	15.3	12
commenting on cohesion	13	7.2
commenting on style	11.9	11.1
commenting on content		
commenting on description	11.9	10.6
commenting on interpretation	13.8	9.7
commenting on extension	11.8	8.3
commenting on form		
commenting on sentence structure	13.6	10.3
commenting on grammatical accuracy	13.9	10.9
commenting on vocabulary and spelling	12.9	12.2

To investigate differences between the two groups in terms of the initial stage of the rating process (i.e., the orientation section in Appendix C) results with respect to the raters' awareness of their performance were examined next. Table 6 shows that the less reliable raters seemed more aware than their more reliable counterparts that they were finding the rating process difficult, with all of them apologising on average at least once (1.3) during the rating of the six scripts and expressing difficulty or inadequacy as they rated every script (6.1) (e.g. 'I think I totally lost the hang of it because it's really hard to mark this'; novice 6). In contrast, the more reliable raters did not apologise at all (0), and on average expressed difficulty or inadequacy in the rating only once (1) during the rating of the six scripts.

Table 6. Awareness of performance: mean frequency of comments by more and less reliable rater groups

Coding category	More reliable	Less reliable
apologising	0	1.3
expressing difficulty or inadequacy rating	1	6.1
commenting on own leniency or harshness	1.7	2.9

Table 7 presents data showing how the two groups engaged with the script during the rating process. In terms of coding, it was the transcriber's notes that largely determined how the coding was done for the categories 'reading script' and 'reading script aloud'. For the most part the term 'reading' or 'reading aloud' was clearly stated in the transcript. Note that the decision on the part of the rater to read scripts aloud or silently was a personal decision in each case; raters were not given instructions as to how or when they were to read scripts (other than that they should read the first sentence of each script aloud so that

researchers would later be able to identify which script was being rated). Careful examination of the transcripts indicated that 'reading' or 'reading aloud' tended to refer to a reading of longer passages of the script. On the other hand, any direct reference to words of the script in the transcript was coded by the researcher as 'citing script'. This tended to be a shorter segment in the context of a specific discussion of some language feature of the script; it is important to note, though, that this also involved a reading aloud of the particular part of the script that was quoted. The respective coding of these two categories thus tended to be discourse driven. Examples of each, taken from the transcription of Novice 8's think-aloud protocol, are given below, with the part of the transcript that was coded for each category in italics:

Reading script

> …Um *(continues reading paragraph)* well, it's hard to interpret exactly what it means, does it mean that you will be isolated or…you will lose your social skills.

Reading script aloud

> um we see a nice argument for watching television together and watching programmes, um *(reads part of last paragraph out loud)* very true, I mean, that's a fairly um it's a fairly complex and nice sociological argument, actually, and I um I tend to agree,

Citing script

> …mentions Sky, um, very nice *'flick of a switch'*, at the end, *'television can also be very relaxing and help people switch off from busy lifestyles'*, very good language, *'however'*, nicely indicated…

The results in Table 7 show that the more reliable raters read the script substantially more, indeed, three times as much (21.4) as the less reliable raters (6.9). However, it is interesting to note that the less reliable raters read the script *aloud* more (12.6) than the more reliable ones did (7.4). This finding perhaps provides some tentative evidence to endorse Barkaoui's (2011) conclusion that reading aloud larger sections of the text hinders global essay comprehension. With regard to citing aloud shorter segments of the script (which according to Barkaoui enables a focus on micro-level problems) the more reliable raters did this over twice (55.7) as much as their less reliable counterparts (21.1). It would not be hard to conclude, then, that rater training would do well to encourage raters to focus more closely on the script. However, these results are in contrast to those obtained by Wolfe et al. (1998), where less proficient (reliable) raters,

using a holistic rating scale, focused more on specific features of the essays they rated. It is interesting to speculate to what extent these differences in results may be a factor of the difference in the type of rating scale used. An analytic scale is more likely to encourage a bottom-up approach to rating (Lievens, 2001), and the results from the present study suggest that focusing on specific features of the text is associated with more reliable rating. On the other hand, this bottom-up approach may not lead to greater reliability when a holistic scale is used (Wolfe et al., 1998).

Table 7. Engagement with script: Mean frequency of comments by more or less reliable rater groups

Coding category	More reliable	Less reliable
reading script	21.4	6.9
reading script aloud	7.4	12.6
citing script	55.7	21.1

Some of the rating behaviours for the two groups are compared in Table 8. Results show that less reliable raters 'introduced a new rater focus' (35.3) and 'referred to the descriptors' (25.7) more than the more reliable raters (19.7/21.1 respectively). The rating category 'introducing new rater focus' was one that was an artefact of the think-aloud protocol, a deliberate indication in the discourse that the rater was changing focus from one rating category in the DELNA descriptors to another (e.g. 'um in terms of *organisation* um it's poorly organised . . ', taken from transcript for Novice 1). This type of discourse move, along with the greater number of references to the descriptors, may be a feature of a lack of familiarity with the DELNA category bands and a need to refer to them more constantly. It would not be hard to imagine that greater knowledge of the categories, and consequently less of a need to both refer to them constantly and to mark this in discourse, would be associated with more reliable rating behaviour.

Table 8. Rating behaviour: Mean frequency of comments by more or less reliable rater groups

Coding category	More reliable	Less reliable
introducing new rater focus	19.7	35.3
referring to descriptors	21.1	25.7
suggesting rating between bands	3.4	2.4

The results in Table 9, which compare the two groups in terms of post-rating behaviour, include not just data from the verbal protocols but also from written comments made as participants read the DELNA benchmarks. It is interesting to note that the more reliable raters were more inclined to 'take on' the DELNA benchmarks by challenging them (2.7) (e.g., 'I don't think that I could go to a 4 though'; Expert 2) and less likely to agree (1) than the less reliable raters (0.9/3

respectively). Conversely, the less reliable raters seemed more aware that there were substantial discrepancies between their rating and the benchmarks (7) (e.g. 'o.k let's see how I did this time – oh crap'; Novice 3) than the more reliable raters (3.4). These results suggest more confidence for the more reliable raters and less self-assurance for the less reliable group. This confidence also helps to explain the fact that the more reliable raters were more likely to point out the problems with the online rater training program (2.7) ('it would actually be helpful if you could put the examples in that, supporting your decisions, so that we can kind of compare on that basis'; Expert 2) than the less reliable raters (1.1).

Table 9. Post rating behaviour: Mean frequency of comments by more or less reliable rater groups

Coding category	More reliable	Less reliable
agreeing with DELNA band	1	3
challenging DELNA band	2.7	0.9
commenting on discrepancy between self and benchmark	3.4	7
commenting on problems with the program	2.7	1.1

Limitations

A number of the limitations of this study have already been presented. We will return again to the issues of the veridicality and reactivity of the think-aloud protocols (Barkaoui, 2011). A re-examination of the think-aloud protocols to look for specific feedback that raters gave revealed that three out of the 6 expert raters reported that they felt that the process of completing think-aloud protocols did have an impact on their rating performance. One spoke about this negatively:

> 'it feels like trying to talk on the phone um and drive a car at the same time, so doing two things at once' (Expert 3)

Another rater mentioned positive aspects of the think-aloud as well as negative:

> 'the stream of consciousness idea did not work very well for me to start with'
>
> 'there's certainly some positives for think aloud, for this protocol um the verbalising, the think aloud protocol forces one to attend to, so specify things and to really try to match the, and understand fully what those descriptors mean' (Expert 4)

Only one novice rater commented on the process of completing a think-aloud. This was in the context of the difficulties it presented.

> 'talking distracts me in fact, it doesn't help me concentrate, I think I find it really difficult' (Novice 3)

The differences in rater backgrounds have already been mentioned as potential limitations in this study. In this regard it is useful, however, to refer to a study by Johnson and Lim (2009). They looked at the influence of rater language background on writing performance assessment and found no pattern of language-related bias in the ratings. The possible impact of differences in educational level and age in the present study is unknown.

Implications and conclusion

This study set out to investigate the possibility of using an online rater training program to train novice raters to rate writing scripts using an analytic scale. The small number of scripts rated was constrained by the inclusion of a verbal protocol in the research design. Results are not conclusive enough to make a case for using this online program for this purpose although there is already some evidence to suggest that it may be an effective tool for 'refresher rater training' for experienced raters. However, as has been previously discussed, a limitation of the present study is that it did not include a measure of the impact of the training on subsequent rating.

This study also undertook an investigation into the rating processes of more and less reliable raters, in an attempt to understand those behaviours that might contribute to more reliable rating. As might be expected, the more reliable raters reported less difficulty in rating and commented less on discrepancies between their ratings and the benchmark ratings than the less reliable raters. Their greater confidence was also demonstrated in their preparedness to challenge the DELNA benchmarks. This finding raises the question of whether it could be useful to ask raters to give an affective response to the rating process as they are experiencing it. The assumption would be that the information obtained may give an indication of how the rater is performing. A predictable outcome of the study was the greater attention paid by the more reliable raters to the DELNA descriptor subcategories and, in particular, their greater engagement with the script. More research is needed to establish the role that focusing on specific features of the text may play in the rating process when an analytic rating scale is used. Results from this study suggest that reading aloud

shorter sections of the text, being consistent with a focus on micro-level text features, was associated with more reliable rating.

The generalisability of this study is limited by its relatively small sample size, that is, 14 raters rating a small number of scripts (six), all written in response to the same prompt. It is hoped that future research will investigate the suitability of online rater training programs where raters are given opportunity to practise rating with a larger number of scripts and where the impact of training can be investigated on subsequent rating, rather than just evaluating the reliability of the ratings generated during the training.

References

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modelling approach. *Language Testing, 27,* 515-535.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28,* 51-75.

Baume, D., Yorke, M. & Coffey, M. (2004). What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment & Evaluation in Higher Education, 29*(4), 451-477.

Brown, A. & Jacquith, P. (2007). *Online rater training: Perceptions and performance.* Barcelona: Language Testing Research Colloquium.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213.

Congdon, P. J. & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.

Dörnyei, Z. (2007). *Research methods in Applied Linguistics: Quantitative, qualitative and mixed methodologies.* Oxford: Oxford University Press.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25,* 155-185.

Elder, C. & Erlam, R. (2001). *Development and Validation of the Diagnostic English Language Needs Assessment (DELNA).* Report prepared for the Vice Chancellor of The University of Auckland.

Elder, C., Knoch, U., Barkhuizen, G. & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly 2, 3*, 175-196.

Elder, C., Barkhuizen, G., Knoch, U. & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*, 37-64.

Ellis, R. & Barkhuizen, G. (2005). *Analyzing learning language.* Oxford: Oxford University Press.

Fox, J. D. (2003). From products to process: An ecological approach to bias detection. *International Journal of Testing, 3*(1), 21-47.

Hamilton, J., Reddel, S. & Spratt, M. (2001). Teachers' perceptions of on-line rater training and monitoring. *System, 29*(4), 505-520.

Johnson, J. & Lim, G. (2009) The influence of rater language background on writing performance assessment. *Language Testing, 26,* 485-505.

Kenyon, D. M., & Stansfield, C. W. (1993). A method for improving tasks on performance-based assessments through field testing. *Language testing: New openings*, 90-102.

Knoch, U., Read, R., & von Randow, J. (2007). Rre-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*, 26-43.

Knoch, U. (2007). *Diagnostic writing assessment: The development and validation of a rating scale.* PhD thesis: The University of Auckland.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255-264.

Loewen, S. & Reinders, H. (2011). *Key concepts in second language acquisition*. New York: Palgrave Macmillan.

Lumley, T. (2000). *The process of the assessment of writing performance: the raters' perspective.* Unpublished PhD thesis: The University of Melbourne.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. New York: Peter Lang.

McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teachers' assessment of ESL writing samples.* Unpublished MA thesis: University of Melbourne.

O'Hagan, S. (2010). *Variability and agreement in assessors' responses to undergraduate essays: an issue for quality assessment in higher education?* Unpublished doctoral thesis. University of Melbourne.

Read, J. (2008) Addressing academic language needs through diagnostic assessment. *Journal of English for Academic Purposes, 7*, 180-190

Pula, J. J. & Huot, B. A. (1993). A model of background influences on holistic raters. In M.M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton.

Robson, C. (2002). *Real world research: a resource for social scientists and practitioner – researchers*. (2nd ed.) Malden, MA: Blackwell.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modelling. *Language Testing, 22,* 1-30.

Schoonen, R., Vergeer, M. & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14,* 157-184.

Stratman, J. & Hamp-Lyons, L. (1994). Reactivity in current think aloud protocols: Issues for research. In P. Smagorinsky (Ed.) *Speaking about writing: Reflections on research methodology* (pp. 89-111) Thousand Oaks, CA: Sage.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11,* 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263-287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6,* 145-178.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (1990). *Communicative language testing.* New York: Prentice Hall.

Wigglesworth, G. (2005). Current approaches to investigating second language learner processes. *Annual Review of Applied Linguistics, 25,* 90-111.

Wolfe, E. W., Kao, C-W. & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15,* 465-492.

Xi, X. & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222-1255.

Appendix A: Writing prompt used for all scripts rated in this study

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.

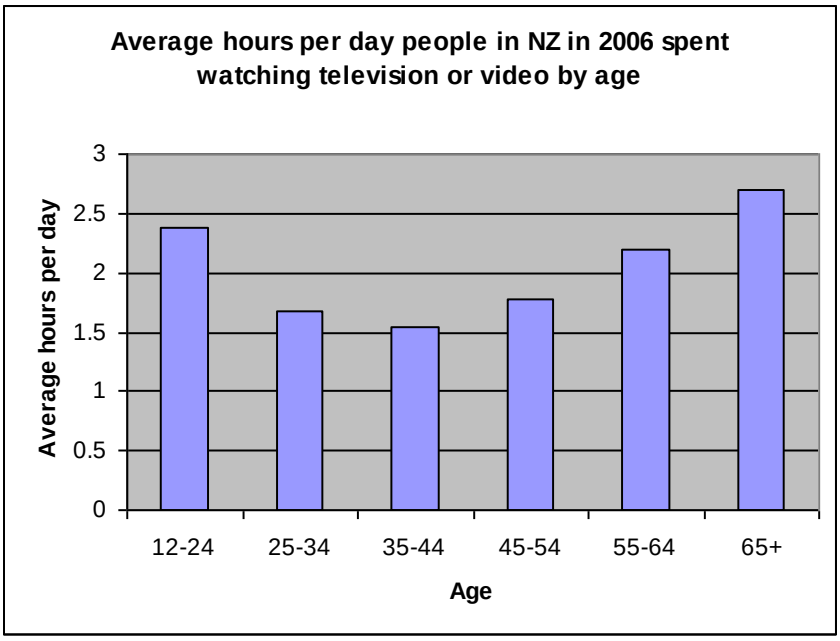
1. Describe the information in the graph.

   THEN

2. Suggest reasons for the trends.

   AND

3. Comment on the benefits and the drawbacks that can result from watching television.



Source: Statistics New Zealand 2006

Appendix B: Screenshot 1

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.

	Your Band	DELNA Band	Discrepancy
Fluency	4	5	1
Content	5	5	0
Form	5	5	0
TOTAL DISCREPANCY			**1**

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?

Appendix C: Comparison of coding by group

Coding category	More reliable	Less reliable
orientation		
commenting negatively	1.1	1.3
commenting on the writer	1.9	1.6
commenting on writer's process	0.1	0.3
commenting positively	1.1	1.3
comparing scripts	1.7	0.1
expressing difficulty or inadequacy rating	6.1	1
comparing with other rating contexts	0.1	0
apologising	0	1.3
engagement with script		
reading script	21.4	6.9
reading script aloud	7.4	12.6
citing script	55.7	21.1
rereading script	0	1.1
pre rating		
describing rating process	6.7	9.3
commenting on own leniency or harshness	1.7	2.9
introducing new rater focus	19.7	35.3
rating categories together	3.3	1.1
commenting on fluency		
commenting on coherence	15.3	12
commenting on cohesion	13	7.2
commenting on style	11.9	11.1
commenting on content		
commenting on description	11.9	10.6
commenting on interpretation	13.8	9.7
commenting on extension	11.8	8.3
commenting on form		
commenting on sentence structure	13.6	10.3
commenting on grammatical accuracy	13.9	10.9
commenting on vocabulary and spelling	12.9	12.2
while-rating		
giving band for particular category	22	24.1
querying difference between bands	0.1	0.7
querying meaning of band	0	1.1
second guessing DELNA Band	0.6	0.1
suggesting rating between bands	3.4	2.4
suggesting a band	15.9	14.7
referring to descriptors	21.1	25.7
justifying reason for a band	10.4	8.6
changing Band	2.3	1.7
post rating		
challenging DELNA band	2.9	0.4
agreeing with DELNA Band	1	3
commenting on discrepancy between self and benchmark	3.4	7
commenting on problems with program	2.7	1.1

Endnotes

[i] For example, fluency scores of 6,5,6 would give an overall score for this category of 5, scores of 6,6,6 would give an overall score for this category of 6.

[ii] In this instance totals are divided by 3 and the band which is closest is allocated. For example, ratings of 5,4,5 for the different categories total 14 which is closer to band 5 than band 4, so the former is given.

[iii] Kappa of 0.40-0.60 – fair; kappa of 0.60-0.75 – good; kappa of above .75 – excellent.