THE VALUE OF RATERS' COMMENTS ON THE WRITING COMPONENT OF A DIAGNOSTIC ASSESSMENT FOR LANGUAGE ADVISING

Stephanie Rummel*

University of Auckland,
Private Bag 92019, Victoria Street West, Auckland 1142, New Zealand

Received 15 March 2020
Revised 20 June 2020; Accepted 22 July 2020

Abstract: The Diagnostic English Language Needs Assessment (DELNA) is used at the University of Auckland to help identify the Academic English needs of students following admission in order to direct them to appropriate support (Elder & Von Randow, 2008). The second tier of DELNA is composed of listening, reading and writing sections, with the writing component rated by trained raters using an analytic rating scale. Language advisers then discuss the marking sheet with the student during an advisory session to provide a detailed overview of the strengths and weaknesses.

The current study was carried out because of difficulties language advisers were experiencing with utilising the marking sheets to draw students' attention to their strengths and weaknesses. A selection of 66 marking sheets with detailed comments from a variety of experienced raters was analysed and coded by two independent researchers. Themes were established regarding features that make a comment valuable or not valuable. Some of those same comments were then shared with students to determine whether or not they agreed with the advisers' assessment. The results show a mismatch at times between language advisers and students. The findings have been used to improve adviser practice and implement a more in-depth rater training programme to help raters better understand the descriptors and to utilise the rating scale to its full potential.

Keywords: Feedback, diagnostic feedback, feedback provision, feedback practices

1. Introduction

Universities in English-speaking countries are increasingly facing challenges as student populations become more linguistically diverse due to growth in the recruitment of international students, immigration inflows and initiatives to broaden participation in higher education by underrepresented groups (Read, 2016). In turn, a growing number of these institutions have begun to rely on post-entry diagnostic language assessments to identify students' academic language needs. According to Lee (2015), the purpose of diagnostics tests is twofold: to identify learners' strengths and weaknesses regarding specific elements of language use and to provide diagnostic feedback linked to remedial learning. These tests often assess students' academic reading, listening and writing skills with the intent of connecting students with resources that can help them appropriately develop in any areas where weaknesses have been identified. Procedures and processes vary among institutions, with the current study investigating the practices

* Tel.: +6493737599 ext 81844
Email: s.rummel@auckland.ac.nz; srummel444@yahoo.com

at the University of Auckland, with a specific focus on the value of comments provided by trained raters on the writing component of DELNA (Diagnostic English Language Needs Assessment), the institution's post-entry diagnostic assessment.

1.1. DELNA at the University of Auckland

DELNA is taken by all first-year students and PhD candidates and is a two-tiered assessment (Read & von Randow, 2016). Students first undertake a computer-based screening that takes about 30 minutes and includes a speedreading activity and an academic vocabulary task. The purpose of the screening is to provide an efficient way to identify proficient users of academic English and exempt them from further assessment (Read, 2008). However, if students fall under a pre-determined cut score, they are required to do a full two hour paper-based diagnosis (two and a half hours if they are a PhD candidate) of their listening, reading and writing skills.

Scores are reported on a scale ranging from 4-9 (Bright & von Randow, 2004). If students receive the highest bands, bands 8 and 9, it is unlikely that they will require academic English language support. Students receiving band 7 may benefit from some support, while band 6 students are thought to need concurrent academic English instruction. However, when a student falls into bands 4 or 5, they are considered at severe risk and in need of urgent language instruction. Those students then attend an advisory session and feedback is provided regarding their results.

1.2 The provision of feedback

According to Hattie and Timperley (2007), the definition of feedback is "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (p. 81). It has an important role in clarifying how well a person is doing and what needs improvement, which enables faster and more effective learning (Hounsell, 2003). Studies have identified various factors that make feedback either helpful or unhelpful. Maclellan (2001) claimed that students may improve their learning when they perceive the feedback to not simply be a judgement of their current level, but as a way to enable learning. Statements that are perceived as being judgemental or unmitigated statements have been found to be unhelpful or lead to defensiveness (Boud, 1995; Hounsell, 1995; Lea & Street, 2000). Weaver (2006) also found that students had difficulty understanding the feedback they received, with a main complaint being that it was too vague to be useful. A further issue identified by her participants was the need to balance negative comments with positive ones so that it would motivate students, which was also identified by Lee (2015) as being important in diagnostic assessments.

In order to be helpful, Lee (2015) posited that diagnostic feedback should establish links between various types of information. Furthermore, the feedback should not only reflect the diagnosis results, but also align itself closely with the resources and learning activities that are available (Lee, 2015). In order to facilitate this, different institutions have implemented varying procedures. Knoch (2012) found that academic advisors played a crucial role in conveying the results to students as they provide human contact in the process. In the case of DELNA, language advisers have delivered students' results since 2005. The position of language adviser was created in response to interview comments from students in which they expressed the desire to receive personalised advice during a one-on-one session (Bright & von Randow, 2004).

DELNA uses the diagnostic assessment to help students reflect on their strengths and weaknesses and a referral form to direct them to appropriate resources that promote academic language development. Any student who receives an average band of 6.5 or lower is asked to attend an advisory session with a DELNA Language Adviser lasting 30-40 minutes for non-PhD students. Any PhD candidate who undertakes the diagnosis attends a one-hour session regardless of their overall band. DELNA language advisers have backgrounds in academic English so they are well placed to help students interpret their results, with positive experiences being reported (Read & von Randow, 2016).

During the consultation, the adviser goes over a language profile that has been generated and includes overall band scores for the three skills that were assessed and computer-generated comments. Then the adviser focusses on the writing and, together with the student, reads through the comments provided by two trained raters regarding the student's writing. The original script is also consulted for specific examples that highlight the strengths and weaknesses. In this way weaknesses are "identified, represented, and described in a detailed and specific manner" (Lee, 2015, p. 304). Knoch (2011) argues that as much detail as possible should be provided from the results of a diagnostic assessment as detailed descriptions of the writer's behaviour allow with tips to improve future performances are more useful.

After various aspects of the writing have been carefully explained, the student is provided with information about workshops and online resources and given a referral sheet in both digital and hard copy to allow easy access. According to the original DELNA principles, there was to be an element of personal choice for students in that although they would be strongly recommended to take advantage of support, they should not be compelled against their will (Read, 2008). However, because questions have arisen regarding whether students actually follow up on recommendations when given the choice (Davies & Elder, 2005; Read, 2013; Knoch, Elder, & Hagan, 2016), currently participation in language enhancement options is required for students at the discretion of their academic programme (Read, 2013). This means that providing a clear description of students' strengths and weaknesses is important because some students may be required to show progress in their language skills before they can progress in their given programme.

1.3 DELNA rating

The quality of the rating is an important consideration in the interpretation of the results of any rater-mediated assessment (Hamp-Lyons, 2007; Johnson, Penny, & Gordon, 2009). In order to ensure validity and reliability, raters must be trained to use the scale to provide detailed feedback on student writing. Training is also important because rater variability may lead to issues such as construct-irrelevant variance (Barrett, 2001; Elder, Knoch, Barkhuizen, & von Randow, 2005; Weigle, 1998). Existing research has focused on rater reliability with issues such as the effectiveness of face to face and online rater training (Weigle, 1998) and rater bias (Weigle, 2011) being investigated, but these have all focussed on matching band scores.

The use of raters' marking sheets during the advisory session means that their comments play an important role in the feedback system utilised at DELNA. As such, on-going training is provided. Because the assessment is diagnostic in nature, it requires a different type of rating scale than those

used for placement and performance, so an analytic scale has been chosen. According to Weigle (2002), analytic scales allow for an indication that different aspects of writing develop at different rates, which provides more useful diagnostic information. Currently, the scale includes nine traits clustered in three categories: coherence and academic style (text organisation, cohesion inside text and academic tone), content (description of data, reasons for trends observed, expansion of ideas), and form (sentence structure, grammatical accuracy, and vocabulary). Each trait is divided into six band levels ranging from four to nine. As raters rate, they are to fill out a marking sheet while referring to graded level descriptors for each trait. There is space on the marking sheet for raters to award a band for each of the nine traits, along with room for them to comment on each trait and provide ticks for correct uses of cohesive devices and referencing. They are also asked to provide crosses for incorrect uses of grammar and vocabulary and language impacting academic style, such as personal pronouns, contractions and informalities. It has been mentioned that some traits might not lend themselves to as fine distinctions as others, which could lead raters struggling to distinguish between the defined levels (North, 2003), so some traits may be more difficult to rate consistently than others.

Because raters' comments are shared with students, for DELNA it is vital that not only the scores match, but also the comments. Furthermore, the comments provide diagnostic information and language advisers must be able to use them to match students' needs with available support, but whether or not comments are valuable to language advisers and what makes a comment valuable have not previously been investigated. According to Kunnan and Jung (2009), "if diagnostic feedback provided to students is not dependable, its practical usefulness is cast into question" (p.617). DELNA language advisers have voiced issues with understanding and using some raters' comments in the past when providing feedback to students and directing them to resources, so the investigation of this issue seemed pertinent so that the training provided to raters could be improved.

2. Materials and Methods

2.1 Aims and research questions

This study aims to improve the comments provided by raters by examining the extent to which language advisers find the comments useful for advising students and students' perceptions of the comments. The research addressed the following questions:

1. What features make a rater's comment on a writing script for a diagnostic assessment valuable for a language adviser during an advisory session with a student?

2. What features reduce the diagnostic value of a rater's comment for a language adviser during an advisory session with a student?

3. To what extent do students' views of the usefulness of specific comments agree with those of the language advisers?

2.2 Methods

The research was carried out in two stages. In the first stage, which took place in 2017 and was used to answer research questions 1 and 2, a selection of 66 marking sheets with detailed comments from a variety of raters with a least two years of experience were chosen at random and analysed and coded by two independent researchers. One researcher was a current DELNA language adviser, while the other had previously been in the same position. Marking sheets were chosen at random to ensure there was a wide range

of comments from different raters. It was decided that 66 sheets would provide a wide range of comments while at the same time allowing themes to emerge. Each marking sheet had raters' comments and band scores for three students on it and for each student there was to be one comment per trait for the nine traits. This means a total of 1,782 comments were analysed. The names of the raters on the rating sheets were covered to ensure anonymity so that the researchers would not be influenced by who had written the comments. The initial codes identified which comments were considered valuable by language advisers in that they allowed the advisers to provide constructive feedback related to specific aspects of students' writing such as grammatical forms, development of ideas, and academic style. The two researchers then worked together and further coding took place to establish themes regarding features such as specificity and clarity that made a comment either valuable or not valuable. This information was entered into a spreadsheet and themes were grouped together. The frequency of a comment being placed into a particular category was also tallied.

In the second stage, which took place in 2019, research question 3 was answered. An email was sent out inviting all students who had completed the diagnosis, received a band score of under 6.5, and been to see a Language Adviser in Semester 1. Five students contacted the DELNA office and all (n=5) were provided with a short survey that included some of the most frequently used comments and they were asked to comment on the usefulness of each. This was followed up with a one-on-one interview (n=4) to gain deeper insight into the students' perspective. Four students were English Language Learners (ELLs) from China, while one was a native speaker of English from New Zealand. One of the Chinese students was a PhD candidate. Of the four Chinese students, three were international students who had been in New Zealand for under a year and one was a permanent New Zealand resident who had been in the country for four years.

3. Results

3.1 Results for research questions 1 and 2

Types of comments that were considered valuable

A two-step process was used to first establish which comments were valuable or not valuable in their professional opinions. See Appendix A for a breakdown of each comment and its categorisation of usefulness. Please note that many comments were made more than once, so for the purpose of this report only each comment is recorded, not the number of times it was made. The researchers then worked together to establish what features made a comment valuable or not. For this step, comments were also checked against the other information on the marking sheets (band number and ticks and crosses) to identify any other issues that may have impacted the value of the comment.

A total of 83.73% (n=1492) of comments examined by the researchers were found to be valuable. The comments that were categorised as most valuable were clear and specific and closely mirrored the descriptors in the analytical scale. In those cases, it was very easy for the Language Adviser to understand why the rater had chosen the band, enabling the Adviser to direct students to appropriate resources. It was also helpful when raters provided information about both strengths and weaknesses that the student exhibited for a particular band. Examples of this were 'paragraphs exist,

but topic sentences unclear' and 'splintered paragraphing, but some organisation of ideas'. The researchers found such comments provided both the Adviser and the student with valuable information about not only what they needed to improve, but also what they were doing well.

Consistency between the bands, the comments and the ticks/crosses was also valuable. It was helpful when the number given by the rater matched the comment provided, for example when a rater said there was some evidence of academic style, a phrase from the band 6 descriptor, and then in turn awarded band 6. In this case, Language Advisers could easily point out to students the areas where they needed improvement.

Another important point was that raters provided a clear comment for each of the nine categories. On the marking sheet, traits are given in the following order: (1) coherence, cohesion, and style; (2) content part 1, part 2, and part 3; (3) sentence structure, grammar, and vocabulary. It was helpful when raters commented in the order of the descriptors, making it clear which trait they were commenting on. Furthermore, when raters included examples in their comments, it was most valuable when they limited the number of examples provided to those that really highlighted the point they were making. Examples of informalities and correct and incorrect use of cohesive devices were particularly helpful because they were clear even when taken out of context.

Types of comments that were not considered valuable

The researchers found that 16.27% (n=290) of comments were not valuable (See Table 1 for specific details). The majority of issues centred around various inconsistencies with the raters' use of descriptor wording (n=145). The most common problem noticed by both researchers was that the comment matched a different band than the one given (n=102). One common example was related to academic style. To receive band 7, the descriptor states the writing should have "most aspects of academic style", for band 6 "some evidence of academic style" and for band 5 "little understanding of academic style". One rater commented that the writing showed "little sense of academic style", but then awarded band 6. At other times, the rater mixed wording from two or more descriptors or two or more traits. In one example, the rater gave band 8; however, the comment said "visible paragraphs, message clear, variable topics, shortish". The wording from this comment matches descriptors from bands 5 (shortish), 6 (variable topics), 7 (visible paragraphs), and 8 (message clear), so it was unclear why an 8 was given.

Other consistency issues were noted to a lesser degree. Raters sometimes double penalised students by, for example, marking them down in both style and vocabulary for informal language. There were also instances when raters penalised students in the wrong place. In the marking sheet there are three headings for comments: coherence/style, content, and form. An example of penalising students in the wrong place may be mentioning grammar errors under coherence/ style rather than form and providing students with a lower band score as a result. Another issue arose when the ticks and crosses given by the rater did not match the comment (n=26). This issue was common in the form categories, where raters often commented that there were numerous grammar errors, but only provided one or two crosses across the categories.

Table 1: Categories of comments that were not valuable

Category	Frequency
Comment does not match band given	102
Examples listed with no context	29
Comment unclear/vague	48
Comment does not match ticks/crosses	26
No comment written	21
Mixed traits described in one comment	21
Comment under wrong trait	14
Difficult to read (handwriting, too much detail)	11
Harsh	10
Double penalisation	8

Both researchers found that some of the comments were unclear. In some cases, they simply did not make sense to the researchers (n=28). One such comment was "organisation is non-academic (has mixed parts)". Both researchers agreed that they were unclear as to what the rater meant. There were also times when the comments used very vague language (n=20) so the researchers were unable to discern the specific problem the rater had identified in the writing, for example "six paragraphs used".

Another issue impacting clarity was the quantity of information given. Some raters provided very detailed comments that became difficult to read given the limited amount of space provided. Others did not write comments for certain categories, often when ticks or crosses had been provided to show correct uses or errors. There were further cases when the raters simply provided lists of words as examples without context so the researchers could not decipher whether the students had used the examples correctly or incorrectly without consulting the original script.

The researchers found a few comments (n=10) that were not constructive as they seemed overly harsh or used too much jargon. Examples of this type of comment include "two topic sentences are non-sensical" and "reasons defy reason!"

3.2. Results for research question 3

In order to answer research question 3, student participants were provided with 17 comments that had been used often in the marking sheets that had been analysed in stage 1 to determine whether or not they found them useful. Most were comments that were found valuable by the language advisers, but a few were ones they thought were not valuable. Table 2 presents the comments language advisers found valuable and Table 3 presents those they felt were not valuable. Each table also includes how many students (n=5) agreed with the language advisers.

Table 2: Number of students who agreed with advisers that comments were valuable

Comment	Number of students who agreed (n=5)
Paragraphs exist, but topic sentences unclearParagraphs exist, but topic sentences unclear	4
Splintered paragraphing, but some organisation of ideas	2

Some paragraphs, but ideas lack organisation and there is repetition as well so it is hard to follow	4
Reasons are clear and well supported with logical development	5
Reasons are inadequate	3
Two reasons provided with adequate support	3
Good use of cohesive devices and clear referencing	5
Overuse of formulaic cohesive devices and repetitious referencing	4
Linking words used well to connect ideas	3
Occasional faulty reference	2
Inadequate range of vocabulary	5
A range of significant grammar errors	3
Article use requires attention	2

Table 3: Number of students who agreed with advisers that comments were not valuable

Comment	Number of students who agreed (n=5)
Organisation is non-academic (has mixed parts	2
Not quite visual paragraphs	3
Goes into substantial waffle about something off the topic	1
Walk/walked, their/there, are/was	2

Students were also asked to comment on why they found a comment valuable or not valuable. In general, when students found a comment to not be valuable, it was because they either did not understand it, or they wanted more specific information to help them understand it. For this reason, comments such as 'splintered paragraphing, but some organisation of ideas', 'occasional faulty reference', and 'article use requires attention' were found to be more valuable to language advisers than to students. The comment with the greatest difference was 'goes into substantial waffle about something off the topic'. Language advisers felt the comment was not valuable because it seemed a bit harsh and they worried that students would not know what was meant by 'waffle'. Students, however, found the comment to be valuable. When asked to explain what the comment meant, most focused on the second part of the comment, and understood they had written something unrelated. The native speaker of English understood the word 'waffle', and did not find it harsh. In the interview she said

> Um, I feel like a lot of lecturers mentioned the last point, about waffle, like don't feel as though you have to write a hundred pages 'cause it means you'll just waffle and completely miss the essay question, which is quite helpful for me…

Besides being given the comments, students were also asked in the interview whether seeing ticks and crosses was helpful. In response, the ELLs all felt it was helpful, with one stating "I think it will be better to get more specific example". However, the native speaker said: "It's not really nice seeing crosses, like what you didn't do. Um, more like maybe constructive feedback, like for next time do this…or you could have done this 'cause Xs can be quite off putting for some people."

4. Discussion

The findings from research question 1 and 2 of this study have implications for rater training in situations where raters are required to provide comments for feedback purposes. Because advisory sessions have been found to play a vital and helpful role in providing students with diagnostic information about their writing (Knoch, 2012; Schuh, 2008; Read & von Randow, 2016), it is important for raters to provide comments that the Language Advisers find useful. Traditional rater training often focuses on band scores; however, in instances when the assessment is diagnostic, comments are equally important as they can be used to better direct students to resources to work on identified weaknesses.

In the case of DELNA, the findings informed an expanded rater training programme for DELNA raters. In 2018, raters were provided with examples of valuable comments and comments that were not valuable and the trainer explained some of the factors that raters should consider when writing their comments. Emphasis was placed on the importance of writing comments that that were clear to they language advisers so that they could explain the comments to the students in language that would be accessible to them, even if they had low levels of language proficiency. Raters' attention was also drawn to key words in the different descriptors that highlight the differences between the bands, because the distinctions between them may not have previously been clear to raters (North, 2003). Furthermore, as most of the raters have experience as either teachers or IELTS examiners, the differences between the type of rating or grading they do in those situations and the type of feedback required for diagnostic assessments was also provided. After initial feedback from raters after the 2018 session, the 2019 training session was further expanded and returning raters were provided with some sample comments that were identified as not valuable and asked to categorise the comments under headings (for example: vague, harsh, etc). A discussion was also had regarding how the comments were used in the advisory session. It was hoped such activities raised raters' awareness so they have a better idea of how their comments are used and the ways they could be improved.

Some of the non-valuable comments were found in a limited number of marking sheets, suggesting they were provided by the same one or two raters. However, other issues such as a mismatch between the comment and the band were more universal. It would therefore seem pertinent to address those widespread problems in depth during the rater training with exercises that allow raters to become more familiar with the band descriptors. Issues that arose in only a few marking sheets could be mentioned during the training, but after rating begins if non-valuable comments are identified as coming from a specific rater, further feedback could be provided in an email.

Of all the identified issues, the frequency of raters awarding a band that did not match the comment is particularly worrying and has been brought to the raters' attention. Inter-rater reliability at DELNA is ensured by matching the marking sheets of two raters. However, only the band awarded is generally considered because there was an assumption that the band and the comment would match. In cases where the band and comment do not match, issues can arise during the advisory session if comments are conflicting, but have been given the same which information to provide to students, which can reduce the face validity of the assessment and also impact the advice being given.

Through raising raters' awareness and sharing experiences of when advisers meet students face to face, it is hoped that raters will give more thought to their comments. This is particularly true regarding the finding that comments that highlight both strengths and weaknesses are valuable, along with the findings that overly harsh comments are not helpful. Alderson and Huhta (2011) point out that diagnostic tests, due to their nature, have a greater focus on weaknesses than strengths. As such, most raters tend to focus on the negative aspects of the writing, but this may be demoralising for some students and that is not the purpose of the assessment. Because some faculties require students to complete a programme after meeting with the Language Adviser (Read, 2013), that they leave their session feeling positive and motivated to engage with the resources available to overcome their weaknesses in academic English is vital. Furthermore, according to Lee (2015), it is desirable to provide learners with information about their weaknesses in parallel with that of their strengths because, for an intervention on weaknesses to be successful, it needs to build on existing knowledge and skills that have already reached or neared the expected level. In this way, weaknesses and strengths may interact and impact the way a learner uses resources provided to enhance areas that have been identified as requiring improvement. The analytical feature of the DELNA scale was designed to allow for this because each criterion should be judged independently.

The findings have also started a discussion regarding the clarity of some of the items on the analytical scale and possible changes that may be made to the rating sheet. DELNA discussed the possibility of designing a rating sheet where raters highlight the relevant parts of the descriptors rather than write their own comments, which would eliminate issues with clarity, mixed descriptors and wrong choice of bands. However, there is a worry that important individualised diagnostic information could be lost if this decision is made. For that reason, it was decided to first provide more in-depth training regarding the comments to see if that would improve the results and raise raters' awareness.

Regarding research question 3, while there was agreement on the value of many, there was disagreement on others. Where there was disagreement, it was often because the student was unclear what the comment meant. This is why the language adviser role is important in the diagnostic feedback process. These comments were provided out of context; however, during the session, language advisers ask questions to try to ensure students understand. They also look through the student's script with them to point out specific examples related to the comments. Because the advisers are professionals in the field of academic writing, they are well placed to provide more explanation during the session and ensure students gain a better understanding of areas needing improvement.

The difference in the response of the native speaker to ticks and crosses is also interesting. As DELNA is administered to the entire student population, regardless of language background, it is important to be sensitive to how native speakers may view receiving feedback on their academic writing. They may also not be very aware of their weaknesses. DELNA seems to be slightly unique from other PELAs in that it is administered to the entire student population, regardless of language background. From experience, many ELLs enter the session with an awareness that their grammar and sentence structure may need some work, but often native speakers do not. Perhaps in those

cases it is best to not focus so much on the crosses highlighting their errors and instead focus more on specific examples in the text that illustrate the point. This is already done in the language advising sessions, but by first showing some students the incorrect use of language in the form of crosses, they may be defensive before reviewing the script with the adviser. The same is true for comments that may be harsh. The goal during the session is to encourage students to use the resources available to improve any identified weaknesses, so it is important that it is not demotivating. However, it is difficult for language advisers to determine beforehand what students may deem as harsh, so language advisers need to be tuned in to students' responses and agile enough to make changes to the session so it suits each individual.

A limitation of the study is the small sample of student participants, so further recruitment could be done to provide a better representation of the student voice. Furthermore, the study could be expanded by investigating the issue from the raters' perspectives. Questionnaires or interviews with raters could be useful in determining reasons for the comments provided and allow for valuable information regarding raters' clarity surrounding the band descriptors. In addition, interviews or reflective journals from language advisers could provide better insight into reactions to the comments and the usefulness of various comments during advisory sessions.

5. Conclusions

The current study identified which comments provided by raters on a diagnostic writing assessment were deemed either valuable or not valuable. Although a robust body of research exists on rater reliability due to its impact on test validity and reliability, studies have mainly focused on test scores. The current study provides important insight into the type of rater training required when raters are asked to provide comments on the writing. In the past, the rater training focussed primarily on the band scores and ensuring raters had the same overall band; however, the findings from the current study emphasise the importance of providing raters with more guidance regarding comments when assessments are used for diagnostic purposes.

The language advisers are in a position to provide individualised feedback to each student who makes an appointment. The process is effective because they not only use the quantitative data contained in the score and the computer-generated comments provided on the profile, but also the qualitative data contained in raters' comments. When valuable comments are provided, they can enrich the advisory session and guide advisers to recommend appropriate resources for academic language enrichment; however, when the comments are not valuable, the adviser needs to spend extra time consulting the script and may even need to skip certain comments during the session. This is difficult during the busy period at the beginning of each semester when back to back appointments leave limited time for such preparation. The better understanding that raters have of how their comments are used and what is considered valuable, the better advisers can direct students. Therefore, enhanced training that goes beyond the band scores should lead to greater benefits for students.

References

Alderson, J.C., & Huhta, A. (2011). Can research into the diagnostic testing of reading in a second or foreign language contribute to SLA research? In L. Roberts, G. Pallotti and C. Bettoni (eds). *EUROSLA Yearbook 11*. John Benjamins, pp. 30-52.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*(1), 49-58.

Boud, D. (1995). Assessment and learning: contradictory or complementary. *Assessment for learning in higher education*, 35-48.

Bright, C., & Von Randow, J. (2004). Tracking language test consequences: The student perspective. Paper presented at the IDP Australian International Education Conference, Sydney. Available online http://aiec.idp.com/uploads/pdf/thur%20-%20Bright%20&%20Randow.pdf

Davies, A. & C. Elder (2005). Validity and validation in language testing. In E. Hinkel (ed.), *Handbook of research on second language learning*. Mahwah, NJ: Erlbaum, 795–813.

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

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, *12*(1), 1–9. https://doi.org/10.1016/j.asw.2007.05.002

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hounsell, D. (2003). Student feedback, learning and development. *Higher education and the lifecourse*, 67-78.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: The Guilford Press.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from?. *Assessing Writing*, *16*(2), 81-96.

Knoch, U. (2012). At the intersection of language assessment and academic advising: Communicating results of a large-scale diagnostic academic English writing assessment to students and other stakeholders. *Papers in Language Testing and Assessment*, *1*(1), 31-49.

Knoch, U., Elder, C., & O'Hagan, S. (2016). Examining the validity of a post-entry screening tool embedded in a specific policy context. In *Post-admission Language Assessment of University Students* (pp. 23-42). Springer International Publishing.

Kunnan, A. J., & Jang, E. E. (2009). Diagnostic feedback in language assessment. *The handbook of language teaching*, 610-627.

Lea, M., & Street, B. V. (2000). Student writing and staff feedback in higher education. *Student writing in higher education: New contexts*, 32-46.

Lee, Y. (2015). Future of diagnostic language assessment. *Language Testing, 32*(3), 295-298. doi:http://dx.doi.org.ezproxy.auckland.ac.nz/10.1177/0265532214565385

Maclellan, E. (2001). Assessment for learning: the differing perceptions of tutors and students. *Assessment & Evaluation in Higher Education*, *26*(4), 307-318.

North, B. (2003). Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. *TOEFL Monograph*, 24.

Read, J. (2008). Identifying academic language needs through diagnostic assessment. *Journal of English for academic purposes*, *7*(3), 180-190.

Read, J. (2013). Issues in post-entry language assessment in English-medium universities. *Language Teaching, 48* (2), pp.1-18.

Read, J. (2016). Post-admission language assessment in universities: International perspectives. *Switzerland: Springer International Publishing*.

Read, J., & von Randow, J. (2013). A university post-entry English language assessment: Charting the changes. *IJES, International Journal of English Studies, 13*(2), 89-110.

Read, J., & von Randow, J. (2016). Extending Post-Entry Assessment to the Doctoral Level: New Challenges and Opportunities. In *Post-admission Language Assessment of University Students* (pp. 137-156). Springer, Cham.

Reinders, H. (2008). The what, why, and how of language advising. In: *MexTESOL*, *32*(2).

Schuh, J. H. (2008). Assessing student learning. In V. N. Gordon, W. R. Habley & T. J. Grites (Eds.), *Academic Advising: A comprehensive handbook*. San Francisco: Jossey-Boss.

Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education*, *31*(3), 379-394.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

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

Weigle, S. C. (2011). Validation of automated scores of TOEFL iBT® tasks against nontest indicators of writing ability. *ETS Research Report Series*, *2011*(2).

GIÁ TRỊ NHỮNG NHẬN XÉT CỦA GIÁM KHẢO CHẤM VIẾT TRONG BÀI THI CHẨN ĐOÁN NHU CẦU TIẾNG ANH

Stephanie Rummel

Trường Đại học Auckland,
Private Bag 92019, Victoria Street West, Auckland 1142, New Zealand

Tóm tắt: Bài thi chẩn đoán nhu cầu tiếng Anh (DELNA) được sử dụng tại trường Đại học Auckland nhằm xác định nhu cầu về tiếng Anh học thuật của sinh viên sau khi nhập học; qua đó, bài thi sẽ giúp nhà trường cung cấp cho sinh viên những hỗ trợ phù hợp nhất (Elder & Von Randow, 2008). Bài thi DELNA hạng hai bao gồm kỹ năng nghe, đọc và viết. Trong đó, bài thi viết sẽ được các giám khảo chấm theo thang chấm phân tích. Các chuyên gia tư vấn ngôn ngữ sau đó sẽ thảo luận phiếu chấm cùng sinh viên trong các buổi tư vấn để mang tới cho sinh viên một cái nhìn tổng quan chi tiết về những điểm mạnh và điểm yếu của các em.

Nghiên cứu này được thực hiện khi các chuyên gia tư vấn ngôn ngữ gặp phải những khó khăn trong quá trình sử dụng phiếu chấm để làm việc cùng sinh viên. Nghiên cứu đã thu thập 66 phiếu chấm với những nhận xét chi tiết từ các giám khảo chấm viết dày dặn kinh nghiệm. Sau đó, hai nhà nghiên cứu độc lập đã tiến hành phân tích và mã hóa các phiếu chấm này. Nghiên cứu đã xác lập được các chủ đề liên quan đến những đặc điểm để đánh giá giá trị của một nhận xét. Một vài nhận xét giống nhau sau đó được gửi tới cho sinh viên để các em quyết định đồng ý hay không đồng ý với những đánh giá của các chuyên gia. Kết quả nghiên cứu cho thấy đôi khi có sự không đồng thuận giữa sinh viên và chuyên gia tư vấn. Những kết quả này đã được sử dụng để cải thiện hoạt động của các chuyên gia và tiến hành một chương trình đào tạo chuyên sâu hơn để giúp các giám khảo chấm viết hiểu rõ hơn về thang chấm và nhờ đó, sử dụng thang chấm hiệu quả nhất.

Từ khóa: phản hồi, phản hồi chẩn đoán, cung cấp phản hồi, hoạt động phản hồi

Appendix A: Raters' comments and whether they were valuable or not

Traits	Comment	Valuable	Not valuable
Coherence	Somewhat random paragraphing		✓
	Some organisation. No visual paragraphs	✓	
	Paragraphing clear. There is an introduction + topic sentences	✓	✓
	Two topic sentences are non-sensical		
	Paragraphing exists as do topic sentences. Message generally clear	✓	✓
	Organised in paragraphs but often needs re-reading		
	Visual paragraphs exist, but places content of some should be in others	✓	✓
	Reasons defy reason!		✓
	Paragraphs exist although a few too many. An introduction and a conclusion exist, but the former is a description, the latter is an irrelevance related to the internet in general	✓	✓
	Visual paragraphs present, but discussion poorly organised with data absent from part 1 but scattered across parts 2 and 3. No clear opening for Part 3	✓	✓
	Some paragraphs but ideas lack organisation and there is repetition as well. Hard to follow	✓	
	Includes some paragraphs but quite waffly and repetitive. Hard to follow. Possibly memorised		✓
	Includes paragraphs- message can generally be followed		✓
	Has used word to show introduction, but essay lacks paragraphs	✓	
	organisation is non academic (has mixed parts)		✓
	Paragraphs used for 3 parts, but few cohesive devices		✓
	Visible paras; messages clear; variable ts, shortish	✓	
	Opening/closing vague		✓
	Splintered paragraphs, short script. Breaks up part 2	✓	
	no visible paras; weak topics; some re reading		✓
	Introduction too general. Paragraphs used effectively to address parts of prompt	✓	
	Has paragraphs but they aren't esp helpful		✓
	Not quite visual paragraphs		
	Intro not very clearly developed/ ideas disconnected	✓	✓
	Ideas not always in logical order		
	Some organisation, some paragraphing. However some parts of the writing require rereading	✓	✓
	Lacks intro statement, only 2 paragraphs, poor org		
	Confused introduction		✓
	Inadequate introductory statement. Has 2 paras but p2 overly long, needs re-reading		
	Some reliance on rubric language		

Cohesion	Cohesion and reference are unnoticeable	✓	✓
	adequate CDs with some referencing		
	Cohesive devices are seemingly simple or not quite	✓	
	accurately used		✓
	Some incorrect CDs		
	CDs: this, third, least most common, as or these due to, not just… but also SIMPLE	✓	✓
	Referencing: One main topic sentence found		
	Some incorrect use of referencing. Formulaic simple		✓
	cohesive devices are used.	✓	
	Some overuse of referencing such as they		
	No referencing. Not many CDs	✓	
	Appropriate range of CDs with good referencing		
	Some good use of CDs. Some used repeatedly. More referencing needed	✓	✓
Style	Chatty lexis interspersed with over formal phrases like	✓	✓
	‘Proof of the above statement is shown or ‘can be obtained		
	by’.	✓	✓
	Hedging adequate		
	Style is appropriate although there is no hedging		
	style is sometimes informal, and often simplistic. Hedging exists	✓	✓
	Many non-academic features: brackets for alternative		✓
	grammatical structures; personal pronouns; chatty vocab	✓	
	Hedging exists		✓
	Informality: the more we rely, rely too much mostly		
	informality centres on direct address through pronouns	✓	✓
	Personal pronouns: who they		
	Little understanding of academic style. Some p/p and		
	rhetorical tone- “it tells us we should”	✓	✓
	Obscure/inconsistent logic		
	Some evidence of academic style- some noticeable		
	wordiness	✓	✓
	rhetoric- more and more		
	Maintains formal register		✓
	Formal but with many errors	✓	
	Little understanding of academic style. Too wordy/informal		
	No actual problems apart from form		✓
	Maintains formal tone + flow of logic. Prose not	✓	
	consistently intelligible. Wordiness		✓
	Maintains academic distance but lacks analysis		
	Some empty sentences		
	Little understanding of academic style, spoken conversational lang, 1st person pronouns x 7, colloquialisms		

Content	No NZ, or 2013, and little data, but overall statements are correct	✓	
	two trends mentioned but briefly		
	One +ve only is inferred		
	Comprehensive	✓	
	No NZ, no 2013, no figures, although trends accurate		
	data description includes place, but not time, and significant figures and trends	✓	✓
	Interpretation is adequate and ideas are relevant with some support		✓
	Time and place given as well as some significant figures(but one figure was misread or wrongly written down) and no mention of figures for train or bicycle	✓	✓
	Interpretation brief	✓	✓
	Ideas generally relevant		
	Interpretation is brief with some irrelevance	✓	✓
	Ideas generally relevant with some support		
	Interpretation is generally adequate and ideas are not always clear	✓	✓
	Part 3 addressed	✓	✓
	Introduction is present, data and trends scattered through essay		
	Paragraph 3 has content repeated from the middle of second paragraph	✓	✓
	Mostly travel in cars…then walked vs. most	✓	✓
	Some relevant ideas but they are not always relevant and lack support		✓
	Along with our health rate decreases	✓	
	Goes into substantial waffle about something off topic	✓	
	Lacks trends but includes figures		✓
	Some reasons are based on assumptions that need substantiating and proof	✓	
	Reason tangential- too much detail on an example	✓	✓
	Lacks overall trends, includes a run down of all figures		
	Some irrelevant reasons and assumptions	✓	✓
	Tangential answer-focused off topic		
	Description includes figures but lacks an overview		
	Lacks clarity- 2 figures 1 mode	✓	✓
	Some reasons for transport lack reason (catching a bus)		
	Gives place and year, notes data comes from a survey; gives main stats and trend	✓	✓
	Combines trends with reasons, environment; price of bikes and availability of bike racks	✓	✓
	Ideas not relevant enough		✓
	Convenience; proximity to work; more busses=fewer trains (x); not tightly structured	✓	
	Partially described- general trends only		
	Very brief and inaccurate reasons	✓	
	Generally adequate		
	Facebook data ok, linkedIn not so detailed. Trends could be more detailed		

Sentence Structure	Numerous fragments on page 1. Page 2, where there are	✓	✓
	complete sentences are grammatically accurate if unvaried	✓	✓
	Adequate if convoluted	✓	✓
	A variety of sentence types, mostly accurate. Punctuation errors	✓	✓
	A variety of sentence types, but rambling and some minor inaccuracies	✓	
	Some rambling and the continuous used for simple several times		
	Some convoluted sentences and punctuation errors		
	Frequent errors in sentence structures	✓	
	range of errors throughout such as punctuation, s+v agreement, sing/pl forms	✓	
	Punctuation: then walked or jogged, third most/then walked or jogged. Third most; , the least/The least	✓	
	Many incomplete sentences and wordiness which make script quite hard to read	✓	✓
	Complex forms contain errors- omissions or incomplete		
	Really wordy. Frequent errors in complex forms		✓
	Word order sometimes off, but most sentences are acceptable with just adequate range	✓	
	frag x1; a few awkward passages; most sentences correctly structured with some variety		
	Very sloppy sentence structure		
	Controlled and varied structures		
Grammar	Minor agreement and article errors	✓	✓
	Minor errors with verbs	✓	✓
	Some minor problems with grammar, especially sing/plural		✓
	Significant basic grammar errors and frequent vocab errors	✓	✓
	Limited control of sentence structure- incomplete and convoluted sentences	✓	✓
	Some basic grammar errors	✓	✓
	Some significant basic errors- articles and misplaced overuse of prepositions		
	On/in, too/to, go/went	✓	✓
	A few minor repeated errors- articles		
	Some repetition/sub + verb agreement? Collocations (higher parking fees)	✓	✓
	repetitious sub+verb agreement		
	People driving (no past tense) s+v agreement/word choice/ expression of ideas is awkward (buses trav. Ten or twentty minutes once) encourage people to catching bus		
	preps incorrect or missing; voice/tense errors; occasional missing word. Confuses be/do		
	missing article; possibly/le; tense; added ‘which’		
	Walk/walked, their/there, are/was		

Vocabulary	Vocab accurate though lacks range	✓	✓
	Simple	✓	
	Vocabulary narrow and repetitive, and some oddities	✓	✓
	Vocabulary accurate but unvaried and a little imprecise	✓	
	Lexically unsophisticated	✓	
	A few wrong choices of vocab but generally appropriate		
	Range and use of vocab inadequate	✓	
	Many borderline vocab choices		
	Vocab is generally appropriate- limited range	✓	
	Range and use of vocab inappropriate- hard to understand	✓	
	Vocab adequate but not always sophisticated		✓
	A few spelling errors but generally appropriate vocab. Limited range	✓	
	Careful but shallow	✓	✓
	Some good vocabulary used, but limited range with grammar structures		