Diagnostic assessment of writing: A comparison of two rating scales
Ute Knoch
Language Testing 2009 26: 275
DOI: 10.1177/0265532208101008

The online version of this article can be found at:
http://ltj.sagepub.com/content/26/2/275
Diagnostic assessment of writing: A comparison of two rating scales

Ute Knoch  University of Melbourne, Australia

Alderson (2005) suggests that diagnostic tests should identify strengths and weaknesses in learners’ use of language and focus on specific elements rather than global abilities. However, rating scales used in performance assessment have been repeatedly criticized for being imprecise and therefore often resulting in holistic marking by raters (Weigle, 2002). The aim of this study is to compare two rating scales for writing in an EAP context; one ‘a priori’ developed scale with less specific descriptors of the kind commonly used in proficiency tests and one empirically developed scale with detailed level descriptors. The validation process involved 10 trained raters applying both sets of descriptors to the rating of 100 writing scripts yielded from a large-scale diagnostic assessment administered to both native and non-native speakers of English at a large university. A quantitative comparison of rater behaviour was undertaken using FACETS. Questionnaires and interviews were administered to elicit the raters’ perceptions of the efficacy of the two types of scales. The results indicate that rater reliability was substantially higher and that raters were able to better distinguish between different aspects of writing when the more detailed descriptors were used. Rater feedback also showed a preference for the more detailed scale. The findings are discussed in terms of their implications for rater training and rating scale development.

Keywords: diagnostic writing assessment, second language writing, second language writing assessment, rating scale development, rating scale validation, rating scales

Alderson (2005) argues that diagnostic tests are often confused with placement or proficiency tests. He lists several specific features which distinguish diagnostic tests from other types of tests. Among these, he writes that diagnostic tests should be designed to identify strengths and weaknesses in the learner’s knowledge and use of language and that diagnostic tests usually focus on specific rather than global abilities.
When discussing the diagnostic assessment of writing, Alderson (2005) describes the use of indirect tests (in this case the DIALANG test) rather than the use of performance tests. However, indirect tests are used less and less to assess writing ability in the current era of performance testing because they are not considered to be adequate and valid measures of the multi-faceted nature of writing (Weigle, 2002) and therefore an argument can be made that diagnostic tests of writing should be direct rather than indirect.

The question, then, is how direct diagnostic tests of writing should differ from proficiency or placement tests. One central aspect in the performance assessment of writing is the rating scale. McNamara (2002) and Turner (2000), for example, have argued that the rating scale (and the way raters interpret the rating scale) represents the de-facto test construct. Accordingly, careful attention needs to be paid not only to the formulation of a rating scale but also to the manner in which it is used. This is all the more important in diagnostic contexts, where it is incumbent upon raters to provide valid, reliable and detailed feedback on the features of learner performance that require further work.

I Rating scales

Several classifications of rating scales have been proposed in the literature. The most commonly cited categorization is that of holistic and analytic scales (Hamp-Lyons, 1991; Weigle, 2002). Weigle summarizes the differences between these two scales in terms of six qualities of test usefulness (p. 121), showing that analytic scales are generally accepted to result in higher reliability, have higher construct validity for second language writers, but are time-consuming to construct and therefore expensive. Because analytic scales measure writing on several different aspects, better diagnostic information can be expected.

Another possible classification of rating scales represents the way the scales are constructed. Fulcher (2003) distinguishes between two main approaches to scale development: intuitive methods or empirical methods. Intuitively developed scales are developed based on existing scales or what scale developers think might be common features at various levels of proficiency. Typical examples of these scales are the FSI family of scales. In recent years, a number of researchers have proposed that scales should be developed based

Rating scales commonly used in the assessment of writing have been criticized for a number of reasons. The first criticism is that they are usually intuitively designed and therefore often do not closely enough represent the features of candidate discourse. Furthermore, Brindley (1998) and others have pointed out that the criteria often use impressionistic terminology which is open to subjective interpretations (Upshur & Turner, 1995; Watson Todd et al., 2004). The band levels have moreover been criticized for often using relativistic wording to differentiate between levels (Mickan, 2003), rather than offering precise and detailed descriptions of the nature of performance at each level.

The problems with intuitively developed rating scales described above might affect the raters’ ability to make fine-grained distinctions between different traits on a rating scale. This might result in important diagnostic information being lost. Similarly, if raters resort to letting an overall, global impression guide their ratings, even when using an analytic rating scale, the resulting scoring profile would be less useful to candidates. It is therefore doubtful whether intuitively developed rating scales are suitable in a diagnostic context.

II The current study

The purpose of this study was to establish whether an empirically developed rating scale for writing assessment with band descriptors based on discourse analytic measures would result in more valid and reliable ratings for a diagnostic context than a rating scale typical of proficiency testing.

The study was conducted in two main phases. During the first phase, the analysis phase, 600 DELNA writing scripts at five proficiency levels were analysed using a range of discourse analytic measures. These discourse analytic measures were selected because they were able to distinguish between writing scripts at different proficiency levels and because they represented a range of aspects of writing. Based on the findings in Phase 1, a new rating scale was developed.
During the second phase of this study, the validation phase, 10 raters rated 100 writing scripts using first the existing descriptors and then the new rating scale. Afterwards, detailed interviews were conducted with seven of the ten raters to elicit their opinions of the efficacy of the two scales.

This paper reports on the findings from the second phase.

Because both qualitative and quantitative data were collected to support the findings, this study is situated in the paradigm of mixed methods research (Creswell & Plano Clark, 2007). More specifically, an embedded mixed methods research model was chosen, where qualitative data are used to supplement quantitative data.

The overarching research question for the whole study is as follows:

**To what extent is an empirically developed rating scale of academic writing with level descriptors based on discourse analytic measures more valid and useful for diagnostic writing assessment than an existing rating scale?**

To guide the data collection and analysis of Phase 2, two more specific research questions were formulated:

**Research question 1:** Do the ratings produced using the two rating scales differ in terms of (a) the discrimination between candidates, (b) rater spread and agreement, (c) variability in the ratings and (e) what the different traits measure?

**Research question 2:** What are raters’ perceptions of the two different rating scales for writing?

### III Method

#### 1 Context of the research

DELNA (Diagnostic English Language Needs Assessment) is a university-funded procedure designed to identify the English language needs of undergraduate students following their admission to the University of Auckland, so that the most appropriate language support can be offered (Elder, 2003; Elder & Von Randow, 2008). The assessment includes a screening component which is made up of a speed-reading and a vocabulary task. This is used to quickly eliminate highly proficient users of English and exempt these from the time consuming and resource-intensive diagnostic procedure.
The diagnostic component comprises objectively scored reading and listening tasks and a subjectively scored writing task.

The results of the DELNA assessment are not only made available to students, but also to their academic departments as well as tutors working in the English Language Self-Access Centre, the Student Learning Centre and on English as a second language credit courses. Based on their results on DELNA, students will be asked to attend language tutorials set up within their specific disciplines, take ESOL credit courses, see tutors in the English Language Self-Access Centre, the Student Learning Centre or take a specific writing course designed for English-speaking background students.

The writing section of the DELNA assessment is an expository writing task in which students are given a table or graph of information which they are asked to describe and interpret. Candidates are given a time limit of 30 minutes. The writing task is routinely double-marked using an analytic rating scale. The scale is described in more detail below.

2 Instruments

a The rating scales: The DELNA rating scale: The existing DELNA rating scale is an analytic rating scale with nine traits (Organization, Coherence, Style, Data description, Interpretation, Development of ideas, Sentence structure, Grammatical accuracy, Vocabulary & Spelling) each consisting of six band levels ranging from four to nine. The scale reflects common practice in language testing in that the descriptors are graded using adjectives like ‘adequate’, ‘sufficient’ or ‘severe’. In some trait scales, different features of writing are conflated into one category (e.g. the vocabulary and spelling scale). An abridged version of the DELNA scale can be found in Appendix 1.

The new scale: The new scale was developed based on an analysis of 600 DELNA writing samples. The scripts were analyzed using discourse analytic measures in the following categories (for a more detailed description of each measure refer to Knoch, 2007a):

- accuracy (percentage error-free t-units);
- fluency (number of self-corrections as measured by cross-outs);
- complexity (number of words from Academic Wordlist);
- style (number of hedging devices – see Knoch (2008));
Diagnostic assessment of writing

- paragraphing (number of logical paragraphs from five paragraph model);
- content (number of ideas and supporting ideas);
- cohesion (types of linking devices; number of anaphoric pronouns ‘this/these’);
- coherence (based on topical structure analysis – see Knoch (2007c)).

To ground the selection of the discourse-analytic measures in theory, several possible models were reviewed as part of the larger study (Knoch, 2007a). These included: models of communicative competence (e.g. Bachman, 1990; Bachman & Palmer, 1996), models of rater decision-making (e.g. Cumming et al., 2001) and models of writing (e.g. Grabe & Kaplan, 1996). As none of these were found to be satisfactory by themselves, a taxonomy of all were used to select the constructs to include in the scale. Measures to represent these constructs were then chosen based on a requirement to fulfil the following criteria: measures had to (a) be able to discriminate successfully between different levels of writing, (b) be practical in the rating process, and (c) occur in most writing samples.

The new scale differs from the existing DELNA scale in that it provides more explicit descriptors. Where possible, raters are given features of writing which they can count. For example, for accuracy, raters are required to estimate the percentage of error-free sentences. The new scale does not make use of any adverbials in the level descriptors, nor does it ask raters to focus on more than one aspect of writing in one trait scale. For a detailed account of the development of the new scale, please refer to Knoch (2007a; 2007b; 2007c). An abridged version of the scale can be found in Appendix 2.

In addition to the qualitative differences of the level descriptors of the two rating scales described above, the two scales differ in the number of band levels of the trait scales. The DELNA scale has the same number of levels for each trait, whilst the new scale has varying number of levels for different traits. The reason for this is that the scale was developed on an empirical basis. The number of levels reflects the findings of the empirical investigation. A comparison of the number of band levels of the two trait scales can be seen in Table 1.

b The writing samples: The one hundred writing scripts used in the second phase of the study were randomly selected from the
scripts produced during the 2004 administration of the DELNA assessment.

c The training manual: To help the raters become familiar with the new scale, a training manual was produced which the raters were asked to study at home before the rater training session. The idea behind the development of this manual was that because the raters were all very familiar with the existing rating scale, a very lengthy training session would have had to be held to introduce them to the new scale. This was, however, not possible because of time constraints on the part of the raters. In the manual, clear instructions are provided on how each trait is to be rated. Each trait scale is further illustrated with examples and practice exercises.

d The interview questions: Because the interviews were semi-structured, the exact interview questions varied from participant to participant. The raters were asked what they thought about the two rating scales, what they would change about the two scales in terms of the wording, categories and number of levels and if they found any categories difficult to apply.

3 Participants

Ten DELNA raters were drawn from a larger pool of raters based on their availability at the time of the study. All raters have several years of experience as DELNA raters and take part in regular training moderation sessions either face-to-face or online (Elder et al., 2007; Knoch et al., 2007).
4 Procedures

a Rater training: The rater training sessions for both the existing rating scale and the new scale were conducted very similarly. In each case, the raters met in plenary for a face-to-face session. In both cases they rated 12 scripts as a group. The raters discussed their own ratings and then compared these to benchmark ratings.

b Data collection: The ratings based on the existing DELNA scale were collected over a period of eight weeks. Two months after the raters had completed their ratings based on the DELNA scale, they rated the same 100 scripts using the new scale. A counterbalanced design was not possible because for practicality reasons the ratings using the existing DELNA scale had to be completed before the new scale was designed. However, because of the large number of scripts in the study, the two months between rating rounds and feedback received from the raters, it can be contended that the raters were not able to remember any of the scripts in the sample from one rating round to the next. To avoid the effect of all raters being previously familiar with the DELNA rating scale, a completely new group of raters could have been recruited for this study. However, as being familiar with the context of the assessment was important for the interviews, and for practicality reasons, this was not done.

After the ratings were completed, all raters were invited to participate in semi-structured interviews. Seven of the ten raters agreed to participate. The interviews were conducted in a quiet room and lasted for 30–45 minutes.

c Data analysis: Three types of data analysis were undertaken: the analysis of the multi-faceted Rasch data, a factor analysis of the rating data and the analysis of the interviews with the raters. Each of these is discussed below.

First, the results of the two rating rounds were analyzed using the multi-faceted Rasch measurement program Facets (Linacre, 2006). FACETS is a generalization of Wright and Masters’ (1982) partial credit model that makes possible the analysis of data from assessments that have more than the traditional two facets associated with multiple-choice tests (i.e. items and examinees). In the many-faceted Rasch model, each facet of the assessment situation (e.g. candidates, raters, trait) is represented by one parameter. The model states that the likelihood of a particular rating on a given
rating scale from a particular rater for a particular student can be predicted mathematically from the proficiency of the student and the severity of the rater.

To interpret the results of the multi-faceted Rasch analysis, a number of hypotheses were developed for comparing the two rating scales:

1) **Discrimination of the rating scale:**

   The first hypothesis was that a well functioning rating scale would result in a high candidate discrimination. When a rating scale is analyzed, the candidate separation ratio is an excellent indicator of the discrimination of the rating scale. The higher the separation ratio, the more discriminating the rating scale is.

2) **Rater separation:**

   The next hypothesis made was that a well functioning rating scale would result in small differences between raters in terms of their leniency and harshness as a group. Therefore, a rating scale resulting in a smaller rater separation ratio is seen to be functioning better.

3) **Rater reliability:**

   The third hypothesis was that a necessary condition for validity of a rating scale, is rater reliability (Davies & Elder, 2005). FACETS provides two measures of rater reliability: (a) the rater point biserial correlation index, which is a measure of how similarly the raters are ranking the candidates and (b) the percentage of exact rater agreement, which indicates, in percentage terms, how many times raters awarded exactly the same score as another rater in the sample. Higher values on both of these indices point to a better-functioning rating scale.

4) **Variation in ratings:**

   Because rating behaviour is directly influenced by the rating scale used, it was further contended that a better functioning rating scale would result in fewer raters rating either inconsistently or overly consistently (by overusing the central categories of the rating scale). The measure indicating variability in raters’ scores is the rater infit mean square value. Rater infit means square values have an expected value of 1 and can range from...
0 to infinity. The closer the calculated value is to 1, the closer the rater’s ratings are to the expected ratings. High infit mean square values (in this case 1.3 was chosen as the cut-off level following McNamara (1996) and Myford and Wolfe (2000)) denote ratings that are further away from the expected ratings than the model predicts. This is a sign that the rater in question is rating inconsistently, showing too much variation. Similarly, low values (.7 was chosen for this study as the lower limit) indicate that the observed ratings are closer to the expected ratings than the Rasch model predicts. This could indicate that a rater is rating very consistently; however it is more likely that the rater concerned is overusing certain categories of the rating scale, normally the inside values.

Each of the four criteria discussed above will be used to compare the two rating scales.

Apart from the multi-faceted Rasch analysis, it was further of interest to ascertain how many different aspects of writing ability the raters were able to discern when using the two rating scales. For this, principal axis factoring was used. This analysis is designed to uncover the latent structure of interrelationships of a set of observed variables. Before this analysis, both the determinant of the R-matrix and the Kaiser-Meyer-Olkin measure of sample adequacy were calculated to ensure suitability of the data to this type of analysis. To determine the number of factors to be retained in the analysis, scree plots and Jolliffe’s (1986) criterion of retaining eigenvalues over .7 were used. Varimax rotation was chosen to make the output of the factor analysis more comprehensible.

The interview data were transcribed and then subjected to a qualitative analysis via a hermeneutic process of reading, analysing and re-reading. The coding themes that emerged during this process were then grouped into categories, including positive and negative comments about each of the two scales.

IV Results

Research question 1: Do the individual trait scales on the two rating scales differ in terms of (a) the discrimination between candidates, (b) rater spread and agreement, (c) variability in the ratings and (d) what the different traits measure?
1 Comparison of individual trait scales

The first step was to compare individual trait scales wherever possible (following Table 1). In the interest of space, only a summary of the findings will be presented with regard to the comparison of individual scales. For the full results please refer to Knoch (2007a).

The findings for the comparison of the individual trait scales generally showed that the trait scales on the new scales resulted in a higher candidate discrimination, smaller differences between raters in terms of leniency and harshness, greater rater reliability, and fewer raters rating with too much or too little variation.

2 Comparison of whole scales

After the individual trait scales were analysed and compared, it was further of interest how the two scales as a whole performed. Figures 1 and 2 present the Wright maps of the two rating scales.

The left-hand column of each Wright map displays the logit values ranging from positive values to negative values. The second column shows the candidates in the sample. Higher ability candidates are plotted higher on the logit scale, whilst lower ability candidates can be found lower on the logit scale. The next column in the Wright map represents the raters. Raters plotted higher on this map are more severe than those plotted lower on the map. Next, the Wright map shows the traits in each rating scale. More difficult traits are plotted higher on the map than easier traits. Finally, the narrow columns on the right of each Wright map represent the trait scales (with band levels) in the order they were entered into FACETS. For the existing scale, these are from left to right: organization (S1), cohesion (S2), style (S3), data description (S4), data interpretation (S5), part three of prompt (S6), sentence structure (S7), grammatical accuracy (S8) and vocabulary/spelling (S9). For the new scale, these are from left to right: accuracy (S1), repair fluency (S2), lexical complexity (S3), paragraphing (S4), hedging (S5), data description (S6), data interpretation (S7), part three of prompt (S8), coherence (S9) and cohesion (S10).

When the two Wright maps were compared, the following observations could be made. First of all, when the raters used the existing scale, the candidates were more spread out, ranging over five logits. When the raters employed the new scale, the candidates were only spread over three logits. Therefore, although most individual trait
<table>
<thead>
<tr>
<th>Measure</th>
<th>Candidate</th>
<th>Rater</th>
<th>Item</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
</tr>
</thead>
<tbody>
<tr>
<td>+4</td>
<td>+9</td>
<td>+9</td>
<td>+9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+3</td>
<td>+9</td>
<td>+9</td>
<td>+9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+2</td>
<td>+***</td>
<td>+9</td>
<td>+9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+1</td>
<td>*****</td>
<td>+9</td>
<td>+9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>***</td>
<td>+9</td>
<td>+9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>*</td>
<td>+9</td>
<td>+9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>*</td>
<td>+9</td>
<td>+9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1** Wright map of DELNA scale
<table>
<thead>
<tr>
<th>Measure</th>
<th>Candidate</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Item</th>
<th>S.1</th>
<th>S.2</th>
<th>S.3</th>
<th>S.4</th>
<th>S.5</th>
<th>S.6</th>
<th>S.7</th>
<th>S.8</th>
<th>S.9</th>
<th>S.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>(9)</td>
<td>(9)</td>
<td>(8)</td>
<td>(9)</td>
<td>(9)</td>
<td>(8)</td>
<td>(8)</td>
<td>(8)</td>
<td>(8)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>*</td>
<td>*</td>
<td>***</td>
<td>6</td>
<td>3</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Repair fluency

Accuracy

Lexical Complexity Part three

Paragraphing

Data

Coherence Interpretation

Cohesion

---

---

---

---

---

---

---

Figure 2  Wright map of new rating
scales on the new scale were more discriminating, it seems that as a whole, the existing scale was more discriminating. This is also confirmed by the first of the rating scale statistics for the whole scale, the candidate separation ratio, displayed in Table 2 below.

It also became apparent that the raters were a lot less spread out when using the new scale. Their severity measures (in logits) ranged from .25 (for the harshest rater) to −.21 (for the most lenient rater), a range of less than half a logit. When employing the existing scale, the raters were spread from .64 to −.74 logits, a range of nearly one and a half logits. That the raters rated more similarly in terms of severity could also be seen by the inter-rater reliability statistics in Table 2, which showed that the exact agreement was higher when the new scale was used (51.2%) than when the existing scale was applied (37.9%). The rater point biserial correlation coefficient, however, was lower when the new scale was used.

Next, the number of raters displaying too much or too little variability in their ratings was scrutinized. For the existing scale, half the raters fell into one of these categories whilst no raters did for the new scale.

When the different traits were examined on the two Wright maps, it became clear that the traits on the new scale were slightly more spread out in terms of difficulty, ranging from .78 on the logit scale for repair fluency to −.74 for cohesion, a difference of one and a half logits. On the DELNA scale the traits spread from .78 (for Data – part three) to −.37 (for style), a difference of just over one logit.

Table 2  Rating scale statistics for entire existing and new rating scales

<table>
<thead>
<tr>
<th>DELNA scale</th>
<th>New scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate discrimination:</td>
<td>Candidate discrimination:</td>
</tr>
<tr>
<td>Candidate separation ratio:</td>
<td>Candidate separation ratio: 5.34</td>
</tr>
<tr>
<td>8.15</td>
<td>5.34</td>
</tr>
<tr>
<td>Rater separation and reliability:</td>
<td>Rater separation ratio: 4.19</td>
</tr>
<tr>
<td>Rater separation ratio:</td>
<td>Rater point biserial:</td>
</tr>
<tr>
<td>8.67</td>
<td>0.38</td>
</tr>
<tr>
<td>Exact agreement: 37.9%</td>
<td>Exact agreement: 51.2%</td>
</tr>
<tr>
<td>Variation in ratings:</td>
<td>Variation in ratings:</td>
</tr>
<tr>
<td>% Raters infit high: 40%</td>
<td>% Raters infit high: 0%</td>
</tr>
<tr>
<td>% Rater infit low: 10%</td>
<td>% Rater infit low: 0%</td>
</tr>
<tr>
<td>Trait statistics:</td>
<td>Trait statistics:</td>
</tr>
<tr>
<td>Spread of trait measures:</td>
<td>Spread of trait measures:</td>
</tr>
<tr>
<td>0.78 to −0.37</td>
<td>0.53 to −0.76</td>
</tr>
<tr>
<td>Trait separation: 9.14</td>
<td>Trait separation: 12.47</td>
</tr>
<tr>
<td>Trait fit values: data and part three much over 1.3, no low</td>
<td>Trait fit values: repair fluency and data slightly high, lexis and coherence low</td>
</tr>
</tbody>
</table>
In a criterion-referenced situation as was the case when these rating scales were used, it is not necessarily a problem to have a bunching up of traits around the zero logit point, as is found in Figure 1 with the traits on the existing rating scale. However, it indicates that raters had difficulty distinguishing between the different traits or that the traits were related or dependent on each other (Carol Myford, personal communication). The fact that the traits in Figure 2 (new scale) were more spread out shows that the different traits were measuring different aspects.

If the traits were not measuring the same underlying construct, then this explains why both the candidate separation of the new scale and the rater point biserial of the new scale were lower than that of the existing scale.

Because the results above only indicate that the traits were measuring different underlying abilities, but not how many different groups of traits the data was measuring, a principal axis factor analysis (or principal factor analysis – PFA) was performed on the rating data. PFA reduces the data in hand into a number of components, each with an eigenvalue representing the amount of variance of the components. Components with low eigenvalues are discarded from the analysis, as they are not seen to be contributing enough to the overall variance. Table 3 (DELNA scale) and Table 4 (new scale) below show the results from the principal factor analysis.

Both the scree plots and the tables displaying the results from the PFA show that when the existing rating scale was analyzed, only one major component was found. This component had an eigenvalue of 5.8 and accounted for about 64% of the entire variance. All other eigenvalues were clearly below 1 (following Kaiser, 1960) and below .7 (following Jolliffe, 1986) and there was no further leveling off point on the scree plot. When the new scale was analyzed,

<table>
<thead>
<tr>
<th>Component</th>
<th>Eigenvalue</th>
<th>% of variance</th>
<th>Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.803</td>
<td>64.472</td>
<td>64.472</td>
</tr>
<tr>
<td>2</td>
<td>.694</td>
<td>7.983</td>
<td>72.455</td>
</tr>
<tr>
<td>3</td>
<td>.657</td>
<td>7.691</td>
<td>80.146</td>
</tr>
<tr>
<td>4</td>
<td>.549</td>
<td>6.341</td>
<td>86.487</td>
</tr>
<tr>
<td>5</td>
<td>.415</td>
<td>4.756</td>
<td>91.243</td>
</tr>
<tr>
<td>6</td>
<td>.275</td>
<td>3.076</td>
<td>94.319</td>
</tr>
<tr>
<td>7</td>
<td>.209</td>
<td>2.326</td>
<td>96.645</td>
</tr>
<tr>
<td>8</td>
<td>.168</td>
<td>1.827</td>
<td>98.472</td>
</tr>
<tr>
<td>9</td>
<td>.138</td>
<td>1.528</td>
<td>100.000</td>
</tr>
</tbody>
</table>

Table 3  Principal factor analysis: Existing DELNA scale
however, the results were different. The PFA resulted in six components with eigenvalues over 0.7.

The next step in the PFA was to identify which variables load onto which component. For this, a rotation of the data was necessary. However, because only one component was identified for the existing scale, no factor loadings can be displayed. A varimax rotation was chosen to facilitate the interpretation of the factors of the new scale. A trait was considered to be loading on a factor if the loading was higher than .4 (as indicated in bold font). The six factor loadings for the new scale can be seen in Table 5.

The largest factor, accounting for 34% of the variance, was made up of accuracy, lexical complexity, coherence and cohesion. This factor can be described as a general writing ability factor. The second factor, which accounted for a further 13% of the variance, was made up of hedging and interpretation of data. This is, at first glance, an

---

**Table 4** Principal factor analysis: New scale

<table>
<thead>
<tr>
<th>Component</th>
<th>Eigenvalue</th>
<th>% of variance</th>
<th>Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.434</td>
<td>34.34</td>
<td>34.34</td>
</tr>
<tr>
<td>2</td>
<td>1.276</td>
<td>12.76</td>
<td>47.10</td>
</tr>
<tr>
<td>3</td>
<td>1.154</td>
<td>11.54</td>
<td>58.64</td>
</tr>
<tr>
<td>4</td>
<td>0.863</td>
<td>8.63</td>
<td>67.28</td>
</tr>
<tr>
<td>5</td>
<td>0.817</td>
<td>8.17</td>
<td>75.44</td>
</tr>
<tr>
<td>6</td>
<td>0.763</td>
<td>7.63</td>
<td>83.07</td>
</tr>
<tr>
<td>7</td>
<td>0.577</td>
<td>5.77</td>
<td>88.85</td>
</tr>
<tr>
<td>8</td>
<td>0.491</td>
<td>4.91</td>
<td>93.75</td>
</tr>
<tr>
<td>9</td>
<td>0.389</td>
<td>3.89</td>
<td>97.64</td>
</tr>
<tr>
<td>10</td>
<td>0.236</td>
<td>2.36</td>
<td>100.000</td>
</tr>
</tbody>
</table>

---

![Scree plots of principal factor analysis](image-url)
unusual factor. However, it can be argued that writers need to make use of hedging in the section where the data is interpreted since the writer is speculating rather than stating facts. For this reason, a writer who scored high on hedging might also have put forward more ideas in this section of the essay. The third factor, which accounted for 12% of the variance, consisted of Part Three of the content, the section in which writers are required to extend their ideas. The fourth factor, which accounted for 9% of the variance, was another content factor, the description of data. That all three parts of the content load on separate factors shows that they were all measuring different aspects of content. Repair fluency was the only measure that loaded on the fifth factor, which accounted for another 8% of the variance. The last factor, which also accounted for 8% of the variance, only had paragraphing loading on it. The six factors together accounted for 83% of the entire variance of the score, whilst the single factor found in the analysis of the existing rating scale only accounted for 64% of the data.

It can therefore be argued that the ratings based on the new scale not only accounted for more aspects of writing ability, but it also accounted for a larger amount of variation of the scores. In other words, there was less unaccounted variance when the new scale was used.

Research question 2: What are raters’ perceptions of the two different rating scales for writing?

The most commonly emerging themes in the interviews were grouped into the following sections:
themes emerging about DELNA scale
themes emerging about new scale

1 Themes emerging about DELNA scale

The most regularly emerging theme in the interviews was that raters often experienced problems when using the DELNA scale. One of the most commonly mentioned problems was that the raters thought the descriptors were often too vague to arrive easily at a score. In the extract below, for example, Rater 4 talked about the problems she encountered when deciding on a score for Content:

Rater 4: [...] And here relevant and supported, I find that tricky support, what exactly is support. Because sometimes it is actually, sometimes you have a number of ideas but there is not much support for them and what is sufficient. [...] You just can’t, there is nothing specific there to hang things on.

Problems with the vagueness of the DELNA descriptors were also reflected in the comments by Rater 5 below:

Rater 5: [...] Sometimes I look at it [the descriptors] I’m going ‘what do you mean by that?’ [...] You just kind of have to find a way around it cause it’s not really descriptive enough, yeah.

A number of raters pointed directly to the adjectives used as being the problem. In the example below, Rater 10 talked about the descriptors for vocabulary and spelling:

Rater 10: Well there’s always a bit of a problem with vocabulary and spelling anyway in deciding you know the difference between extensive, appropriate, adequate, limited and inadequate. So there’s sort of adverbial [sic]. Yeah, it’s really just a sort of adverbial thingy anyway isn’t it so I think I just go with gut instinct on that one.

Although most raters reported having problems deciding on band levels with the DELNA scale, the methods of coping were quite different for different raters. A variety of strategies (both conscious and subconscious) emerged from the interviews. These were as follows:

- assigning a global score
- rating with a halo effect
- disregarding the DELNA descriptors.

The first strategy that a number of raters referred to in their interviews was assigning a global score to a script, usually after the first reading.

Rater 5 below describes his rating process, which is more holistic than analytic:
Rater 5: Mmh, yeah, I always automatically think, this is a native speaker, this is a non-native speaker. How well will this come across, will it be sufficient for academic writing and then that is sort of borderline between six and seven quite often and then is it a better seven or is it an eight or is it less than a six, or is it five.

This overall, holistic type rating often results in a halo effect, where a rater awards the same score for a number of categories on the scale. Below, Rater 10 talked about awarding scores for the three categories grouped under fluency in the DELNA scale, organization, cohesion and style:

Rater 10: For style, again, I just tend to go with the gut instinct. And I suspect I often tend to give the same grade or similar grades for cohesion and style. Probably for the whole of fluency. [...] So in a way, it is almost like giving a global mark for the three things in consideration. With, if someone had no paragraphing, but everything else was good, maybe a bit of variation.

Some raters seemed to clearly disregard the DELNA descriptors and override them with their own general impression of a script. Rater 10 (below) was talking about the score she would award for organization to a script that had no clear paragraph breaks but was otherwise well organized. The DELNA descriptors recommend awarding either a five or a six.

Rater 10: Mmh [...] well, I think I would, (sighs), looking at this it ought to be a six, but it is possible particularly if I suspected that it was a native speaker, and that it was someone that wasn’t so strong in academic writing but actually had very good English, I might even go up to a seven, but I [...] yeah, if I had other reservations about the language and stuff, then I would give it a six or even a five if it is really bad. But if I was sort of convinced by the writer in every other way, I might well push the score up in a way not to pull them down. Just for the paragraphing.

2 Themes emerging about new scale

The most commonly emerging theme about the new scale was that the raters liked the fact that the descriptors in the new scale were more explicit. This is evident in the following extracts from the interviews:

Researcher: Do you feel you used the whole range there [accuracy in the new scale]?

Rater 10: Yes, yes, I did. Yeah. I think I would be more likely to. Because I thought I had something to actually back it up with, it had a clearer guideline for what I was actually doing, so I was more confident for giving nines and fours. And I think also because I didn’t, I let go of the sense of this is a seven, so I have to make it come out as a seven and I’d say, well, sorry, if they have no error-free sentences they get a four and I don’t care if it is something that
might otherwise get a six or a seven and yes, if they can write completely error-free then I can give them a nine. I have no problems with that.

The idea of being able to arrive precisely at a score was also echoed in the following comment by Rater 7:

Rater 7: It is interesting, I found that it [the new scale] is quite different to the DELNA one and it is quite amazing to be able to count things and say, I know exactly which score to use now.

Whilst the comments about the new scale reported above shed a positive light on the scale, a less positive comment was also made by the raters.

Three raters criticized the fact that some information was lost because the descriptors in the new scale were too specific. Rater 5, for example, argued that a simple count of hedging devices could not capture variety and appropriateness:

Researcher: You said that, other than hedging, style wasn’t really considered.

Rater 5: Yeah, it does seem a bit limited. […] I suppose that is similar [to the DELNA scale] it sort of relies on the marker’s knowledge of English in a more kind of global way sort of. But maybe that is the inter-rater reliability issue coming up.

Above, the results for research questions 1 and 2 were presented. The following section aims to discuss these results in light of the overarching research question:

To what extent is an empirically developed rating scale of academic writing with level descriptors based on discourse analytic measures more valid for diagnostic writing assessment than an existing rating scale?

V Discussion

DELNA is a diagnostic assessment system. To establish construct validity for a rating scale used for diagnostic assessment, we need to turn to the limited literature on diagnostic assessment. Alderson (2005) compiled a list of features which distinguish diagnostic tests from other types of tests. Four of Alderson’s 18 statements are central to rating scales and rating scale development. These are shown in Table 6.

This section will discuss each of Alderson’s four statement in turn and then focus on the raters’ perceptions of the two scales.
Statement 1. Diagnostic tests are designed to identify strengths and weaknesses in a learner’s knowledge and use of language.

Alderson’s first statement calls for diagnostic assessments to identify strengths and weaknesses in a learner’s knowledge and use of language. Both rating scales compared in this study were analytic scales and were designed to identify strengths and weaknesses in the learners’ writing ability. However, the PFA showed that the new scale distinguished six different writing factors, whilst the current DELNA scale resulted in one large factor. Therefore, it could be argued that the new scale was more successful in identifying different strengths and weaknesses.

The main reason that the ratings based on the DELNA scale resulted in only one factor was the halo effect displayed by most raters. Although developed as an analytic scale, the existing scale seemed to lend itself to a more holistic approach to rating. It is possible that, as hypothesized in this study, the rating scale descriptors do not offer raters sufficient information on which to base their decisions and so raters resort to a global impression when awarding scores. This then would explain why, when using the empirically developed new scale with its more detailed descriptors, the raters were able to discern distinct aspects of a candidate’s writing ability.

Some studies have in fact found that raters display halo effects only when encountering problems in the rating process (e.g. Lumley, 2002; Vaughan, 1991). Lumley, for example, found that when raters could not identify certain features in the descriptors, they would resort to more global, impressionistic type rating. This study suggests that the halo effect and impressionistic type marking might be more widespread than has so far been reported. It was possible to show that simply providing raters with more explicit scoring criteria can significantly reduce this effect. It could therefore be argued that

---

Table 6  Extract from Alderson’s (2005) features of diagnostic tests

1. Diagnostic tests are designed to identify strengths and weaknesses in a learner’s knowledge and use of language.
2. Diagnostic tests should enable a detailed analysis and report of responses to items or tasks.
3. Diagnostic tests thus give detailed feedback which can be acted upon.
4. Diagnostic tests are more likely to be ... focussed on specific elements than on global abilities.
the halo effect is not necessarily only a rater effect, but also a rating scale effect.

However, what was not established in this study was whether the raters rated analytically because they were unfamiliar with the new scale. It is possible that extended use of the new scale might also result in more holistic rating behavior. Further longitudinal research is needed to determine whether this is indeed the case.

Statement 2. Diagnostic tests should enable a detailed analysis and report of responses to items or tasks and

Statement 3. Diagnostic tests thus give detailed feedback which can be acted upon.

Alderson’s (2005) second and third statements assert that diagnostic assessments should enable a detailed analysis and report of responses to tasks and that this feedback should be in a form that can be acted upon. Both rating scales lend themselves to a detailed report of a candidate’s performance. However, as evident in the quantitative analysis, if the raters at times resort to a holistic impression to guide their marking when using the DELNA scale, this will reduce the amount of detail that can be provided to students. If most scores are, for example, centred around the middle of the scale range, then this information is likely to be less useful to students than if they are presented with a more jagged profile of some higher and some lower scores.

A score report card based on the new scale could be designed to make clearer suggestions to students. For example, the section on academic style could suggest the use of more hedging devices or students could be told how they could improve the coherence of their essays rather than just being told that their writing ‘lacks academic style’ or is ‘incoherent’. More detailed suggestions on what score report cards could look like are beyond the scope of this paper, but can be found in Knoch (2007a).

Statement 4. Diagnostic tests are more likely to be [...] focussed on specific elements than on global abilities.

Alderson’s fourth statement states that diagnostic tests are more likely to be focussed on specific elements rather than on global abilities. If a diagnostic test of writing should focus on specific elements, then this needs to be reflected in the rating scale. Therefore, the descriptors need to lend themselves to isolating more detailed aspects of a writing performance. The descriptors
of the new scale were more focussed on specific elements of writing because they were based on discourse analytic measures which had, in an earlier phase of this research, been found to discriminate between texts at different levels of proficiency. To arrive at a truly diagnostic assessment of writing, all categories in an analytic scale need to be reported back to stakeholders individually, otherwise the diagnostic power of the assessment is lost.

1 Raters’ perceptions of the two scales

Finally, it was also important to establish the raters’ perceptions of the efficacy of the two scales for diagnostic assessment. Raters’ perceptions of the scale usefulness are important as they provide one perspective on the construct validity of the scale. As language experts they are well qualified to judge whether the writing construct is adequately represented by the scale.

In the course of the interviews it became apparent that most raters treated DELNA as a proficiency or placement test rather than a diagnostic assessment. For example, Rater 10 commented: *I think I prefer the existing DELNA scale because I like to mark on ‘gut instinct’ and to feel that a script is ‘probably a six’ etc. It was a little disconcerting with the ‘new’ scale to feel that scores were varying widely for different categories for the same script.* Similarly, Rater 5 mentioned in his interview: *I notice these things [features of form] as I am reading through, but I try not to focus too much on them. I try to go for broad ideas and are they answering the question. Are they communicating to me what they need to communicate first of all. And how well do they do that.* It seems therefore that the diagnostic purpose of the assessment was not clear to them and/or that their experience of rating different kinds of tests, such as IELTS, was influencing their behaviour. The findings of this study suggest that raters need to be made aware of the purpose of the assessment in their training sessions, so that they recognize the importance of rating each aspect of writing separately. This might result in raters displaying less of the halo or central tendency effects.

Returning to the overall purpose of this study, the following conclusions can be drawn. Bachman and Palmer (1996) suggest that ‘the most important consideration in designing and developing a language test is the use for which it was intended’ (p. 17). We need, therefore, to remember that the purpose of this test is to provide
Diagnostic assessment of writing

detailed diagnostic information to the stakeholders on test takers’ writing ability. Most evidence speaks in favour of the new scale. One aspect not reported on previously has to do with practicality. Two aspects of practicality have to be taken into consideration: (1) practicality of scale development and (2) practicality of scale use. It is clear that the scale development process for an empirically developed scale is more laborious. In terms of practicality of use, the new scale proved only minimally more time-consuming. Some raters even reported being able to use the new scale faster. Overall, it can be argued, however, that the new scale is less practical.

But, as Bachman and Palmer (1996) and Weigle (2002) argue, it is impossible to maximise all aspects of test usefulness. The task of the test developer is to determine an appropriate balance among the qualities in a specific situation. Since each context is different, the importance of each quality of test usefulness varies from situation to situation. Test developers should therefore strive to maximize overall usefulness given the constraints of a particular context, rather than try to maximize all qualities. In the context of DELNA, a diagnostic assessment, it could be argued that construct validity is central (as is the case in most assessment situations). Practicality is always a crucial consideration, but wherever possible, construct validity should not be sacrificed simply to ensure practicality.

Overall, the new scale has been shown to generally function more validly and reliably in the diagnostic context that it was trialled in than the pre-existing scale.

VI Conclusion

The findings of this study have a number of implications. The first refers to the classification of rating scale types commonly found in the literature. Weigle (2002), as well as many other authors distinguishes between holistic and analytic rating scales However, this study seems to suggest that although these two types of scales are distinct, it is also necessary to distinguish two types of analytic scales: less detailed, a priori developed scales and more detailed, empirically developed scales. Therefore, Weigle’s summary table can be expanded in the following manner (Table 7).

Researchers and practitioners need to be made aware of the differences between analytic scales and need to be careful when making decisions about the type of scale to use or the development method to adopt.
Another implication relates to score reporting. Table 8 presents different purposes for which writing tests are administered. The table provides a short description of the purpose of each test, what type of rating scale might be used and how the score should be reported. For example, whilst for a proficiency test it might be less important if the rating scale in use is holistic or analytic (as long as it results in reliable ratings), the rating scale used in diagnostic assessment would need to be analytic and at the same time should provide a differentiated score profile. The need for these different types of scales is a consequence of the way the scores are reported. Results of tests

### Table 7 Extension of Weigle’s (2002) table to include empirically developed analytic scales

<table>
<thead>
<tr>
<th>Quality</th>
<th>Holistic scale</th>
<th>Analytic scale – intuitively developed</th>
<th>Analytic scale – empirically developed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>Lower than analytic but still acceptable.</td>
<td>Higher than holistic.</td>
<td>Higher than intuitively developed analytic scales.</td>
</tr>
<tr>
<td>Construct Validity</td>
<td>Holistic scale assumes that all relevant aspects of writing develop at the same rate and can thus be captured in a single score; holistic scores correlate with superficial aspects such as length and handwriting.</td>
<td>Analytic scales more appropriate as different aspects of writing ability develop at different rates. But raters might rate with halo effect.</td>
<td>Higher construct validity as based on real student performance; assumes that different aspects of writing ability develop at different speeds.</td>
</tr>
<tr>
<td>Practicality</td>
<td>Relatively fast and easy.</td>
<td>Time-consuming; expensive.</td>
<td>Time-consuming; most expensive.</td>
</tr>
<tr>
<td>Impact</td>
<td>Single score may mask an uneven writing profile, may be misleading for placement and may not provide enough relevant information for diagnostic purposes.</td>
<td>More scales can provide useful diagnostic information for placement, instruction and diagnosis, but might be used holistically by raters; useful for rater training.</td>
<td>Provides even more diagnostic information than intuitively developed analytic scale; especially useful for rater training.</td>
</tr>
<tr>
<td>Authenticity</td>
<td>White (1985) argues that reading holistically is a more natural process than reading analytically.</td>
<td>Raters may read holistically and adjust analytical scores to match holistic impression.</td>
<td>Raters assess each aspect individually.</td>
</tr>
</tbody>
</table>
Diagnostic assessment of writing

Table 8 Rating scales and score reporting for different types of writing assessment

<table>
<thead>
<tr>
<th>Purpose of writing test</th>
<th>Definition</th>
<th>Rating scale</th>
<th>Score reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proficiency test</td>
<td>Designed to test general writing ability of students.</td>
<td>Holistic or analytic.</td>
<td>One averaged score.</td>
</tr>
<tr>
<td>Diagnostic test</td>
<td>Designed to identify strengths and weaknesses in writing ability; designed to provide detailed feedback which students can act upon; designed to focus on specific rather than global abilities.</td>
<td>Needs to be analytic and needs to result in differentiated scores across traits.</td>
<td>In detail; separate for each trait rating scale.</td>
</tr>
</tbody>
</table>

of writing proficiency are usually only reported as one averaged score but, as Alderson (2005) suggests, the score profiles of diagnostic tests should be as detailed as possible and therefore any averaging of scores is not desirable.

Overall, this study has been able to show that a rating scale with descriptors based on discourse-analytic measures is more valid and useful for diagnostic writing assessment purposes. The uniqueness of diagnostic assessment has recently again been highlighted by Alderson’s (2005) book *Diagnosing foreign language proficiency* and therefore this study is an important contribution to the writing assessment literature. The author has attempted to show that not all analytic rating scales can be assumed to be functioning diagnostically and that scale developers and users need to be careful when selecting or developing a scale for diagnostic assessment.

**VII References**


Diagnostic assessment of writing


## Appendix 1: Abridged DELNA scale

<table>
<thead>
<tr>
<th></th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FLUENCY</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td>Essay fluent – well organised</td>
<td>Evidence of organization</td>
<td>Little organization – possibly no paragraphing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>– logical paragraphing</td>
<td>– paragraphing may not be entirely logical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohesion</td>
<td>Appropriate use of cohesive devices – message able to be followed throughout</td>
<td>Lack / inappropriate use of cohesive devices causes some strain for reader</td>
<td>Cohesive devices absent / inadequate / inappropriate – considerable strain for reader</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Style</td>
<td>Generally academic – may be slight awkwardness</td>
<td>Some understanding of academic style</td>
<td>Style not appropriate to task</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CONTENT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description of data</td>
<td>Data described accurately</td>
<td>Data described adequately / may be overemphasis on figures</td>
<td>Data (partially) described / may be inaccuracies / very brief / inappropriate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpretation of data</td>
<td>Interpretation sufficient / appropriate</td>
<td>Interpretation may be brief / inappropriate</td>
<td>Interpretation often inaccurate / very brief / inappropriate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development / extension of ideas</td>
<td>Ideas sufficient and supported. Some may lack obvious relevance</td>
<td>Ideas may not be expressed clearly or supported appropriately – essay may be short</td>
<td>Few appropriate ideas expressed – inadequate supporting evidence – essay may be short</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FORM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Structure</td>
<td>Controlled and varied sentence structure</td>
<td>Adequate range – errors in complex sentences may be frequent</td>
<td>Limited control of sentence structure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammatical accuracy</td>
<td>No significant errors in syntax</td>
<td>Errors intrusive / may cause problems with expression of ideas</td>
<td>Frequent errors in syntax cause significant strain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary &amp; spelling</td>
<td>Vocab. appropriate / may be few minor spelling errors</td>
<td>Limited, possibly inaccurate / inappropriate vocab / spelling errors</td>
<td>Range and use of vocabulary inadequate. Errors in word formation &amp; spelling cause strain</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Appendix 2: Abridged new scale

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repair</td>
<td>Error-free</td>
<td>Nearly error-free</td>
<td>No self-corrections</td>
<td>No more than 5 self-corrections</td>
<td>Large number of words from academic wordlist (more than 20)</td>
<td>Use of large number of sophisticated words</td>
</tr>
<tr>
<td>Fluency</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Complexity</td>
<td>Error-free</td>
<td>Nearly error-free</td>
<td>No self-corrections</td>
<td>No more than 5 self-corrections</td>
<td>Large number of words from academic wordlist (more than 20)</td>
<td>Use of large number of sophisticated words</td>
</tr>
<tr>
<td>Paragraphting</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Hedges</td>
<td>No hedging devices</td>
<td>1 paragraph</td>
<td>4-8 hedging devices</td>
<td>5-8 hedging devices</td>
<td>More than 9 hedging devices</td>
<td>Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times</td>
</tr>
<tr>
<td>Mechanics</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Reader-Writer Interaction</td>
<td>Error-free</td>
<td>Nearly error-free</td>
<td>No self-corrections</td>
<td>No more than 5 self-corrections</td>
<td>Large number of words from academic wordlist (more than 20)</td>
<td>Use of large number of sophisticated words</td>
</tr>
<tr>
<td>Context</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Data</td>
<td>All data described (all trends and relevant figures)</td>
<td>Most data described (all trends, some figures)</td>
<td>More than 9 hedging devices</td>
<td>5-8 hedging devices</td>
<td>3-4 hedging devices</td>
<td>Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times</td>
</tr>
<tr>
<td>Interpretation of data</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Ideas</td>
<td>Five or more relevant reasons and/or supporting ideas</td>
<td>Four or more relevant ideas</td>
<td>Few or no relevant reasons and/or supporting ideas</td>
<td>Few or no relevant reasons and/or supporting ideas</td>
<td>Few or no relevant reasons and/or supporting ideas</td>
<td>No relevant reasons and/or supporting ideas</td>
</tr>
<tr>
<td>Part 3 of task</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Coherence</td>
<td>Error-free</td>
<td>Nearly error-free</td>
<td>No self-corrections</td>
<td>No more than 5 self-corrections</td>
<td>Large number of words from academic wordlist (more than 20)</td>
<td>Use of large number of sophisticated words</td>
</tr>
<tr>
<td>Writer makes regular use of super structures, sequential progression and possibly indirect progression</td>
<td>Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writer might use this/these to refer to ideas more than four times</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohesion</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

All data described (all trends and relevant figures) | Most data described (all trends, some figures) | All data not attempted or incomprehensible | Data description not attempted or incomprehensible | 1 paragraph | No hedging devices | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times |

No self-corrections | No more than 5 self-corrections | Large number of words from academic wordlist (more than 20) | Use of large number of sophisticated words | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times | 1 paragraph | No hedging devices | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times |

No self-corrections | No more than 5 self-corrections | Large number of words from academic wordlist (more than 20) | Use of large number of sophisticated words | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times | 1 paragraph | No hedging devices | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times |

No self-corrections | No more than 5 self-corrections | Large number of words from academic wordlist (more than 20) | Use of large number of sophisticated words | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times | 1 paragraph | No hedging devices | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times |

No self-corrections | No more than 5 self-corrections | Large number of words from academic wordlist (more than 20) | Use of large number of sophisticated words | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times | 1 paragraph | No hedging devices | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times |

No self-corrections | No more than 5 self-corrections | Large number of words from academic wordlist (more than 20) | Use of large number of sophisticated words | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times | 1 paragraph | No hedging devices | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times |

No self-corrections | No more than 5 self-corrections | Large number of words from academic wordlist (more than 20) | Use of large number of sophisticated words | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times | 1 paragraph | No hedging devices | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times |

No self-corrections | No more than 5 self-corrections | Large number of words from academic wordlist (more than 20) | Use of large number of sophisticated words | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times | 1 paragraph | No hedging devices | Frequent: Unrelated progression, coherence breaks and some extended progression, there is little cohesion. Writer might use this/these to refer to ideas more than four times |