Karien Hattingh and Johann L. van der Walt
North-West University (Potchefstroom Campus)

The development and validation of a rating scale for ESL essay writing

Abstract

This article describes an empirical procedure for developing and validating a rating scale for assessing essays in English as a second language. The study was motivated by a concern for the validity of the scoring grid currently used to assess ESL essay writing at Grade 12 in the final end-of-year examination in South Africa. Following an argument-based validation framework based on the work of Kane, we describe the development, trialling and revision of a rating scale. An empirical procedure, based on an analysis of a sample of Grade 12 ESL essay writing, was followed to develop a new rating scale. The validation process is presented in four phases as part of a specification of an evaluation inference.

Keywords: writing assessment, validation, scoring validity, rating scale development, ESL writing, multi-faceted Rasch measurement

Journal for Language Teaching, Volume 47, Number 1, pp. 73-105. 2013.
ISSN 0259-9570. DOI: http://dx.doi.org/10.4314/jlt.v47i1.4

1. Introduction

There is general agreement that rating scales for writing should be based on actual samples of learner writing (cf. North & Schneider, 1998; Alderson, 1991; Fulcher, 1987, 1993, 2003; Upshur & Turner, 1995; Turner, 2000; Douglas, 2001; Weigle, 2002) as opposed to features which raters imagine are relevant distinguishing indicators at specific performance levels. However, many national school examinations use scales that have been drawn up in a non-empirical fashion by committees consisting of examiners and teachers, and these are accepted as empirically reliable and valid. A typical example is the rating scale used in the Grade 12 school-leaving (matriculation) examination in the public school system in South Africa.

We argue that empirical scale development, as opposed to non-empirical development, yields a more valid assessment instrument, which, together with sufficient rater training and standardised assessment procedures, is key to improving scoring validity (which includes reliability), particularly in high-stakes examinations (cf. Bachman, 1990; Alderson, Clapham & Wall, 1995; Fulcher, 2003; Weir, 2005; Shaw & Weir, 2007).

Relatively few validation studies describe the empirical process followed in the development of a rating scale, and there is little step-by-step guidance for this process (cf. Kane, 2001, 2004). Knoch (2009) describes the validation process of a diagnostic rating scale. Our article contributes to closing this gap by describing the procedure followed in the empirical development and validation of a new rating scale for assessing essay writing in an achievement test, and focuses in particular on the reliability of ratings given in response to the scale. The context is the assessment of Grade 12 English Second Language (ESL) essays in the public South African National Senior Certificate (NSC) school-leaving examination.

2. Framework for the validation of writing assessment

Current conceptions of validity regard it as an argument concerning test interpretation and use (cf. Messick, 1989; Chapelle, 1999, 2012; Kane, 2001, 2006, 2012; Bachman & Palmer, 2010). It concerns the interpretation of test scores rather than the scores themselves. As validity can only be accessed via validation (Davies & Elder, 2005: 796), it is necessary to formulate a validity argument in any validation exercise.

A number of writers have argued that rating scales should have a theoretical basis and that the construct to be measured should be defined in advance (e.g. McNamara, 1996; North, 2003; Weir, 2005), but, as Knoch (2009: 73) points out, there is at present no single theory of writing that can serve as basis for the design of a rating scale. Chapelle, Enright and Jamieson (2010: 4) arrived at a similar conclusion with regard to a validation of the TOEFL – they found it difficult to base their validity argument on a theory of language proficiency, as “no agreement exists concerning a single best way to define constructs of language proficiency to serve as a defensible basis for score interpretation”.

They state that Kane's (e.g. 2001, 2006, 2012) argument-based perspective offers a different perspective to score interpretation, and offers a solution to this problem, making validation more accessible than it has been before. This approach does not require a theory or construct per se, although it does not disregard applied linguistic discussion of language ability constructs completely. Kane's (2006:23) approach requires an explicit statement of the proposed interpretation and uses of scores – an interpretive argument – followed by a validity argument that evaluates the interpretive argument. Kane (2006:23) points out that an interpretive argument lays out "the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performance". It involves the collection of evidence in support of the proposed interpretations. The validity argument involves a critical evaluation of the proposed interpretations, based on quantitative and/or qualitative data. In terms of Toulmin's (2003) argumentation model, rebuttals and counterevidence should also be considered in this argument.

Test score interpretations are always based on inferences. Kane (2001:330) mentions five basic inferences in assessment: evaluation, generalization, extrapolation, explanation and decision-making or utilization. In the assessment of essays, a rater has to arrive at a score after reading the essay. The rater can only do so by means of an inference; in this case, an evaluation inference. Chapelle et al. (2010:10) argue that an evaluation inference should rest on a description of the domain of interest, which they also call an inference. Inferences should be specified in as much detail as possible, and this involves a statement of the warrants, assumptions and backing involved. Any inference is supported by a warrant, which rests on an assumption that in turn requires backing.

The evaluation inference in this article has the warrant that test taker essays are evaluated to provide ratings that reflect Grade 12 ESL writing ability. This warrant rests on the assumption that the criteria in the rating scale are relevant for and critical to scoring ESL essays in the NSC examination and that these are applied correctly and appropriately (cf. Kane, Crooks & Cohen, 1999:9; Chapelle et al., 2010:8). This assumption should be backed by evidence of an iterative empirical process that entails the development, trialling and refinement of the scale. The whole process results in the general claim that the test-takers can communicate effectively in writing in English. Sub-claims may include statements that test-takers can organise ideas coherently, express their own opinions, produce extended pieces of writing, and formulate ideas on a variety of topics.

The development of a rating scale amounts to part of what can be termed the design validity of a writing test. It forms one part of the validity argument (viz. evaluation inference), which is a progressive step-by-step process across "bridges" to a conclusion about test score use (Kane et al., 1999: 9; Chapelle et al., 2008: 9). In describing the empirical process followed in developing and validating a rating scale, this article also demonstrates the types of evidence that can be collected to support the evaluation inference.

3. Background to the study

In terms of the series of inferences specified by Chapelle et al. (2010), the following background to the study forms part of the domain inference. In the final English Second Language matriculation writing paper test-takers usually have a choice between a narrative, descriptive, discursive or an argumentative topic on which to write an essay of 250-300 words. These options are given in order to provide for the wide variety of test taker backgrounds, and are an established tradition in the examination. (For this reason, the essays used in this study were not written on the same topic.) This situation requires a generic rating scale that can accommodate all these genres. The essay counts 50 marks out of a total of 100 for the writing paper. Test-takers are given two hours to complete the writing paper, which, in addition to the essay, also requires them to complete two pieces of transactional writing (e.g. a letter or report and travel directions or an agenda). The marks for the writing paper are added to the marks for the language and literature papers to obtain a final examination mark.

The present scale (Appendix A) for the assessment of essay writing in the Grade 12 examination assesses them in terms of levels ranging from 1 to 7. Each level is linked to a range of marks that can be allocated to it. These ranges are indicated in the relevant blocks where rows and columns meet. Although the essay counts 50 marks, a range of percentages is also indicated next to the code to assist with mark allocation. This scale was originally developed more than a decade ago by a panel of experienced examiners. It was therefore based on expert opinion (cf. Knoch, 2009:14). No empirical data on its scoring validity are available. It contains only two criteria, viz. language and content, which are not clearly distinguished. It is also unlikely that two criteria can provide an adequate representation of a complex construct like writing (cf. Knoch, 2009:73). A rater has to consider a complex variety of features under each criterion, and it is not very easy to distinguish among them. Some descriptors in the scale are also not very clear, e.g. what does “critical awareness of the impact of language” mean, and how is it assessed? It is also not clear whether proof-reading and editing should indeed be assessed, as evidence of these is not stated as a requirement in the question paper. The scale also makes imprecise distinctions such as adequate and moderate. Our experience, supported by feedback from teachers, has been that these are vague and unclear distinctions that cause confusion and result in inconsistent scoring, particularly when used by relatively unskilled and undertrained raters. Marks tend to be bunched around the average, and raters find it difficult to discriminate between performance levels, resulting in good essays often being assigned average marks. These poorly-focused criteria and ill-phrased descriptors seem to contribute to rater variability, slowing down the scoring process and, ultimately, resulting in scores that are not totally reliable (cf. McNamara, 1996: 121; McNamara, 2000: 38; Weir, 2005: 180-198).

4. Data collection and analysis

The development and validation of the new rating scale consisted of four phases: A benchmarking exercise to establish examples of typical learner writing at the various

levels (Phase One); drafting a new scale by a panel of experts, based on data from the first phase (Phase Two); refining it (Phase Three); and piloting the new scale by a panel of typical examiners (Phase Four). These four phases outline the procedure for collecting relevant evidence in order to support the assumptions inherent in the validity argument so as to support or refute the validity claim.

A randomly selected sample of 200 essays written by Grade 12 learners in a final examination was collected to serve as typical examples of learner performances. Based on the mark originally assigned to each essay by examiners (using the current rating scale), the sample was divided into seven levels of writing proficiency, as the department of education uses a seven-point scale as its standard in all school subjects. The essays were coded: Level 1 indicated essays scoring 20% or lower, and Level 7 indicated essays scoring 80% and higher. A working sample of sixty-eight essays was selected from the original 200, which included essays at each level, and excluded problematic performances – e.g. essays that were much too short, or ones in which only the prompt was copied – to ensure a representative sample for further analyses.

Quantitative analyses were conducted during each of the four phases by means of Multi-faceted Rasch Measurement (MFRM) procedures (Linacre, 2006). MFRM is an application of Item Response Theory. It is a logistic latent trait model of probabilities that calibrates different facets independently of each other, within a common frame of reference. All facets are measured on a logit scale. Thus, different facets, viz. test-takers' ability, rater severity and task difficulty, can be compared to one another. Fit statistics give an indication of the degree to which each facet conforms to, or disagrees with, other relevant facets when measuring the trait in question. MFRM expects variance, and accounts for it, but facets that either fit the model too poorly (misfit) or too perfectly (overfit) are considered problematic in terms of the acceptable range of fit statistics.

There are no hard and fast rules for determining what degree of "fit" (i.e. the range of accepted variance) is acceptable (Weigle, 1998: 276). Upper- and lower-control limits may vary (Park, 2005: 9; Coniam, 2010: 428). With 1.0 considered a "perfect fit" (Bond & Fox, 2007: 285-286) (fit values greater than 1.0 pointing to misfit and less than 1.0 to overfit), some researchers suggest a narrow range with a lower control limit of 0.70 or 0.75 and an upper control limit 1.30 (cf. McNamara, 1996; Bond & Fox, 2001; Eckes, 2005). Others, such as Wright and Linacre (1994), Weigle (1998) and Linacre (2002), regard lower and upper control limits of 0.05 and 1.50 respectively as acceptable. As the assessment of writing is not a hard-and-fast science, fit values in the range between 0.5 and 1.5 were considered acceptable for the purposes of this article. MFRM can also provide measurements of degrees of inter-rater consistency, as well as person-item interaction (intra-rater consistency) (McNamara, 1996: 121; Schaefer, 2008: 466). Raw scores alone may be an under- or over-rated view of performance due to different degrees of rater severity (Engelhard, 1992: 98; McNamara, 1996: 118). All Rasch analyses were conducted using the FACETS version of the Multi-faceted Rasch program (Linacre, 2006).

The aim of the analysis in each phase is described in the discussion of each below. Various panels of experts were involved in each phase, and these are also described.

5. Phase One, benchmarking exercise

The aim of Phase One was to benchmark examples of typical learner performances at seven scale levels. This accords with the requirement specified in the South African National Curriculum Statement (South Africa, Department of Education, 2005) that writing should be assessed by means of a seven-point scale, as mentioned above.

A panel of fourteen experienced ESL raters from four provinces scored the sixty-eight essays. They included markers, deputy chief markers, chief markers, and internal moderators used by the department of education in order try and limit discrepancy resulting from lack of experience or undertraining in scoring writing as these are factors known to influence scoring reliability. The original scores assigned to the essays by teachers were not taken into consideration. On average, the panel had nineteen years' experience of marking Grade 12 ESL examination essays. All participants were familiar with the rating scale currently in use, and had a thorough knowledge of the context in which assessment takes place. This rating was unavoidably done with the current scale, but we argued that these raters would be able to provide benchmarked ratings despite any inadequacies in the rating scale because of their levels of expertise and experience. Scoring took place in two intervals of four weeks each. Raters each received a set of unmarked and typed copies of the essays to score at home. Each rater scored at least thirty-two essays, and each essay was scored by at least nine of the fourteen raters.

Results were processed statistically by means of MFRM procedures. In Phase 1, FACETS was used to investigate the following:

- the degree to which the sample of essays represented the full range of abilities on the scale;
- inter-rater consistency;
- criterion (item) difficulty (language and content);
- the accuracy of the levels at which essays were benchmarked, based on raters' scores, and
- the appropriate benchmark level for individual essays.

The Rasch measurement procedure was repeated a total of three times to eliminate all extreme (miss- and/or overfitting) cases. During the first two calibrations, four essays were identified as outliers – they had values greater than 1.5 (misfitting) or smaller than 0.5 (overfitting) – and were removed from the sample. In each of the four cases, considerable disagreement between scores was reported. Data on the remaining sixty-four essays were calibrated for a third time, and no outliers were identified. The logit scale in Figure 1 presents the results of this calibration exercise, mapping the interaction between learner ability (Essays 1-64, second column), rater severity (Raters 1-14, third

column), and item difficulty (Criteria 1-2, language and content, fourth column). The Measure in column one displays the Rasch logit scale and the Scale in column five the seven levels of writing ability.

Essays ranged across 10 logits (-4 to +6) and all seven scale levels (column five), with more essays grouped around the middle than toward the ends of the scale. The distribution of the essays showed that the selected sample of essays was sufficiently representative of the range of performance levels to establish typical examples at different achievement levels. Raters were placed between -1 and +1 logits, with the exception of Rater 2 (below -1), indicating sufficient inter-rater agreement, with some variation as expected.

Both language and content are placed in line next to the 0 logit mark, indicating that these 'items' are not distinguished, i.e. they do not measure different aspects of the construct in question, as argued above. Neither one was consistently scored more harshly than the other, nor was any significant bias towards either criterion reported. (Data on rater bias were obtained from the FACETS analysis, but are not discussed in detail in this article.) No significant infit or outfit mean-square values were therefore reported for the essays, raters or criteria items.

In addition to the vertical ruler report, FACETS reports a reliability index (similar to Cronbach's alpha) that indicates accuracy in distinction (e.g. how accurately raters distinguish between levels of proficiency or scale criteria), with values closer to 1 signifying accurate distinction between factors (cf. Myford & Wolfe, 2003, 2004). In this case a very high reliability index of 0.98 was reported for essays (learner ability), a high value of 0.93 for raters and a below acceptable value of 0.63 for the criteria language and content. The fact that highly experienced and well-trained raters achieved an acceptable level of reliability using this scale should not it itself be considered sufficient evidence to support generalising claims about the not typical of the average rater in the final examination.

```
-------------------------------------------------------------------------
|Measr|+Essay            |-Rater        |-Criteria|Scale|
-------------------------------------------------------------------------
+   6 +                  +              +         + (7) +
|    |                  |              |         |    |
|    |                  |              |         |    |
|    | 2                |              |         |    |
|    |                  |              |         |----|
+   5 + 32               +              +         +    +
|    |                  |              |         |    |
|    |                  |              |         |    |
|    |                  |              |         |    |
|    | 37               |              |         | 6  |
+   4 +                  +              +         +    +
|    | 31               |              |         |    |
|    | 55  60           |              |         |----|
|    |                  |              |         |    |
|    | 14  19  46       |              |         |    |
+   3 +                  +              +         +    +
|    | 20  38  51  63   |              |         | 5  |
|    | 23               |              |         |    |
|    | 34  45  48       |              |         |    |
|    |                  |              |         |----|
```

+ 2 + + + + +

| | 17 35 49 | | | |

| | 12 42 | | | |

| | 13 24 62 | | | 4 |

| | 53 58 | | | |

+ 1 + + + + +

| | | 10 | | |

| | 57 | 11 14 8 | | --- |

| | 15 36 50 52 7 | 4 | | |

| | | 1 5 7 9 | | |

* 0 * 43 * * 1 2 * *

| | | | | 3 |

| | 22 8 | 12 13 | | |

| | 16 29 33 41 44 | 3 | | |

| | 4 | 6 | | --- |

+ -1 + 30 + + + +

| | 11 5 | | | |

| | 54 59 | 2 | | |

| | 61 | | | 2 |

| | 10 27 56 9 | | | |

+ -2 + 26 39 + + + +

| | 21 40 | | | |

| | 47 | | | --- |

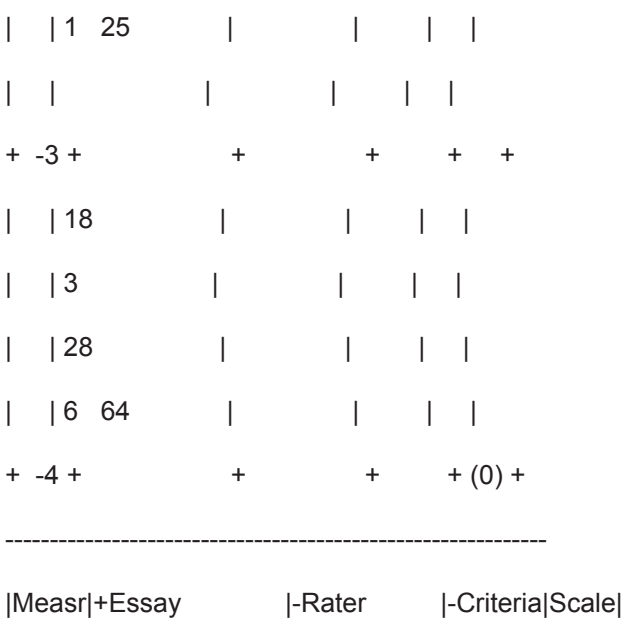

Figure 1: FACETS vertical ruler report for Phase 1 benchmarking exercise

The FACETS vertical ruler report (Figure 1) indicates the estimated true level of ability on the scale (as opposed to the observed values or true scores), with rater variance considered. The placing of essays as indicted in Figure 1 (second column) in terms of ability levels on the scale (fifth column) was therefore used to assign benchmark levels to the sixty-four essays.

After this analysis, the sixty-four essays were established as typical examples of Grade 12 learner writing across the seven performance levels. They were re-numbered for procedures in the following phases.

6. Phase Two, drafting the new rating scale

A draft scale was compiled in Phase 2. In order to achieve this, the salient features of essay writing were identified in the sixty-four benchmarked sample scripts and categorised. Level descriptors and criteria were then formulated at each of the seven proficiency levels.

A new seven-member panel from three different L1 backgrounds (five experienced raters, a departmental external moderator and one of the authors) took part in this phase. The participants were selected on the basis of their extensive experience and expertise in the fields of ESL assessment, L2 writing and scale development. They shared an average of twenty-five years' teaching experience and eighteen years' ESL scoring experience at matriculation and first-year university level.

Each participant received background reading for the project and copies of the numbered and labelled benchmarked essays to analyse prior to a workshop. Individual findings were reported and compared during the two-day workshop.

In their discussion of the most appropriate scale format during the workshop, the panel first consulted the National Curriculum Statement (South Africa, DoE, 2005) for a definition of the construct of writing in question, but the document is very vague in this regard, providing only a number of assessment outcomes to be achieved. The panel also examined the essay questions in the Writing paper. It became apparent that the skill of writing at this level was multi-faceted and that the scale would need to address a number of aspects, such as topic knowledge and insight, organisation, grammar, and sentence construction. In other words, a multi-faceted taxonomy would be needed to ensure comprehensive assessment of Grade 12 essays. The panel also investigated typical, established scales, such as those of Jacobs, Zingraf, Wormuth, Hartfield and Hughey (1981), IELTS (2007) and the TEEP Attribute Writing Scale (Weir, 1990).

Based on these considerations, the panel agreed on an analytic scale, and not a holistic one, as the most appropriate means of assessing essay writing. This decision was made in order to prevent construct over- and/or under-representation. Bachman and Palmer (2010: 324), amongst others, also express their preference for an analytic scale as it allows for a more nuanced scoring and ties the instrument directly to the construct. Individual elements of the construct are stipulated, allowing for clearer delineation of the construct, and therefore more control over whether it is being over- and/or under-represented. Multi-faceted Rasch procedures allow for more specific identification of both threats in terms of fit statistics.

The scale was drafted in the form of a seven-point Likert-type semantic differential scale with extreme bi-polar descriptors. Semantic differential scales provide binary terms (such as "black" or "poor" as opposed to "white" or "excellent") at the ends of a continuum according to which raters evaluate the degree to which a performance accords with these extremes (Hattingh, 2009:187-188). It was argued that bi-polar descriptors provided a specific description of the range into which performances could be categorised and that they eliminated ambiguous interpretation of criteria. A total score would be calculated by adding scores for individual features to a total out of 100.

To establish a suitable assessment taxonomy, the panel conducted an analysis of the writing performances in the essays; first at micro- (individual features that stood out

typifying different performance levels) and then at macro-level (i.e. categories of features). The panel members compared their individual findings and consolidated differences, then analysed the essays once more as a group to check that the taxonomy items identified were indeed relevant and to ensure that all performance-distinguishing features had been identified. Following an emergent-coding approach (cf. Haney, Russell, Gulek & Fierros, 1998; Stemler, 2001, 2004), they identified the most salient features of writing (at micro-level) to be included in the scale as items in the taxonomy, and then arranged these into five categories (at macro-level) to be used as major criteria. The panel finally agreed on fifteen micro-level features as the most prominent distinguishers between different achievement levels, and then grouped related ones together. After the second re-drafting, five macro-categories emerged, viz. Content, Structure and development, Grammar, Vocabulary and Editing.

The number of micro-level features in each macro-category criterion determined the weighting of the particular criterion in the scale. The more features listed under a macro-criterion, the heavier the weighting of that criterion in the scale. Each micro-criterion would be scored individually on a seven-point scale, apart from Item 15 (at this stage Presentation) under the criterion (at this stage still called) Editing, which was allotted two marks. The panel formulated descriptors for the opposing poles of each of the fifteen individual micro-categories. In addition to the draft scale, the panel compiled a scoring guide that clarified the criteria and descriptors in detail.

Four weeks after the workshop, five panel members (excluding the authors) blindly scored thirty unlabelled benchmarked essays, representing all seven performance levels, using the draft scale. They scored individually at home. A calibration exercise was then conducted to establish the scoring consistency among the five raters using the draft scale, rater bias towards any of the criteria, and the degree to which micro-criteria were distinct yet relevant for the assessment in question, i.e. to check the reliability of the selection of the categories statistically. For this exercise, data were subjected to a Rasch analysis. MFRM can accommodate items that are scored on different scales in the same analysis as long as the necessary specifications are stipulated in the input file. Thus it was possible to calibrate the results for the fifteenth micro-criterion − a dichotomous item counting only 2 marks – in the same analysis as results on the other fourteen micro-features, scored on a seven point scale. Despite the difference in weight, the performance of micro-feature 15 could be compared to that of the other micro-features on the same logit scale.

This procedure verified the panel's analyses of salient features of writing, providing support for the proposed criteria and the fifteen items identified as most prominent indicators of performance. The Rasch vertical ruler reported indicated sufficient inter-rater consistency, with all raters placed within 1 logit measure, although a negative placement (between -1 and -2) indicated that all the raters displayed slightly harsh scoring tendencies. All fit measures reported for raters were, however, within the acceptable range of 0.5 – 1.5. Furthermore, the items were scored consistently and could thus be used to distinguish different levels of ability in performances. Rasch reported a high reliability index of 0.95 for the raters. A low index indicates that raters are in agreement (they score "as one"),

and show little inter-rater variability. A high reliability value indicates variance between raters, but this is not problematic – it is expected in practice and the MFRM assumes it.

As far as the individual distinctiveness of the scale criteria and their relevance were concerned, the results were favourable. All micro-criteria, apart from Item15 (at this stage called Presentation), were grouped closely around the 0 logit mark. This could indicate that these individual features tap into related aspects or be interpreted as showing that the raters were awarding very flat profiles across the criteria. However, a high reliability correlation of 0.93 for the items confirmed the apparent distinction between the items in the vertical ruler report. Furthermore, infit- and outfit mean-square measures for the items were all within the set parameters. Item 15, the dichotomous item, was placed as clearly distinct from the other features, which seemed to indicate that this criterion was 'easier' than the others and could be addressing a different aspect. This was not unexpected, as Presentation could be considered a surface feature of writing, rather than an inherent feature such as organisation at sentence and paragraph level, or vocabulary and spelling. It is thus not problematic that this feature was placed separately. The infit and outfit mean-squares reported for Feature 15 were 0.72 and 1.48, which fall within the acceptable range of fit values. These values support the panel's decision not to exclude this micro-criterion (at this stage Presentation) from the scale.

Results from Phase Two Rasch calculations supported the panel's selection of the five macro-criteria and fifteen micro-criteria, verifying the draft scale to be tested and refined in Phase Three.

7. Phase Three, refinement of the new scale

The aim of Phase Three was to refine the draft. A third panel critically evaluated the scale in a series of scoring and discussion sessions to identify potential weaknesses, and revise it accordingly. In addition to the authors, the panel consisted of ten qualified and experienced ESL teachers from different language and cultural backgrounds and teaching environments. They shared an average of twenty-one years' marking experience, and represented a range of schools from well-performing, privileged schools to disadvantaged and underprivileged schools. Both quantitative and qualitative data were collected.

During a two-day workshop the panel trialled the draft scale, subjected it to content analysis, and refined it. The workshop started with a blind-scoring session during which the ten teachers each scored two essays using the draft scale without any discussion of its content.

After the blind scoring, the panel evaluated the criteria and features in the scale in terms of the degree to which they adhered to the specifications in the National Curriculum Statement's Language Programme Guidelines (South Africa, DoE, 2008a) and the Subject Assessment Guidelines (South Africa, DoE, 2008b). In addition, they discussed

the format of the scale, the criteria, and their formulation and organisation in the scale in order to identify any aspects that may interfere with the clarity and consistent interpretation and application of the instrument.

Following this evaluation session, the draft was revised. For this exercise, the panel divided into three groups, which each then proposed solutions to the problems areas identified. Each group presented their solutions as a refined version of the draft. The three revised versions were compared and differences solved through a panel discussion. Once the participants had reached consensus on all problematic aspects, a final version was drafted, containing adapted criteria labels, bi-polar descriptors and organisation of the scale. Changes made to the original draft scale involved the distribution of features to ensure a fairer weighting of criteria, and the revision of bi-polar descriptors to avoid ambiguity by providing explicit and extreme end-scale level labels.

Revising the weighting of criteria resulted in a re-distribution of items included under each of the five macro-criteria and complete revision of the last one, Editing. Both Editing and Presentation remained problematic during the revision. The panel finally decided to replace Presentation (included under Editing) with Length, again weighted only two marks. The argument was that tidy presentation could not be considered a skill indicative of writing proficiency, whereas the question paper stipulated length requirements to which test-takers had to adhere. Their inability/ability to do so should be penalised or rewarded.

Potential ambiguous descriptors were identified and revised, with a focus on providing descriptions that clearly reflected extreme performances at the end of a continuum. For example, the original bi-polar description Demonstrating a lack of insight into and understanding of the topic (Level 1) versus Demonstrating insight into and understanding of the topic (Level 7) was rephrased as No insight into and understanding of topic (Level 1) versus Outstanding insight into and comprehensive understanding of topic (Level 7).

The workshop ended with a calibration exercise in which we used MFRM to investigate the performance of the revised draft scale. The panel scored five ‘clean’ benchmarked essays in this exercise. The data were subjected to a Rasch analysis to investigate the same aspects as in Phase Two and misfitting or overfitting facets were reconsidered.

Significant rater variability was reported for two raters, with infit mean-squares of 2.36 and 1.95, and outfits of 2.51 and 2.12 respectively. The MFRM report also indicated that the first of these raters displayed bias in scoring the first five features of the first essay; the second one showed bias tendencies in scoring the first and seventh micro-features in two essays. In such cases, the relevant datapoints should be deleted from the dataset for further analyses. The reliability index reported for raters was high at 0.95.

Slight misfit was reported for micro-feature 7 at 1.55 infit and 1.58 outfit, and was not regarded as reason for concern. Significantly misfitting and overfitting values of 2.22 infit mean-square and 4.51 outfit mean-square were reported for Item 15 (Length). Generally, such a misplaced item would be rejected, as such a placement may indicate

that the feature is not a particularly good discriminator, or may be too 'easy'. (In this case, the length would either be appropriate, slightly too long/short, or much too long/short) After reconsidering their previous decision concerning this item, the panel reached consensus that it was an essential requirement in the scale. We again concluded that the overfit was due to the allocation of only two marks to this criterion, as opposed to seven marks for the others. Furthermore, a high reliability index of 0.96 was reported for the micro-categories, also indicating that the features were sufficiently distinct and that each micro-criterion addressed an individual aspect of the writing skill.

Directly after scoring, each rater was asked to provide a one-page written feedback report, responding to an open-ended question about their scoring experience and opinion of the scale (in terms of its format, content and organisation of the items, ease of use, clarity and comprehensibility). This feedback supported the quantitative results, i.e. that the scale aided consistent scoring and provided an accurate and relevant taxonomy of items relevant to the assessment of essays. The raters indicated that the scale was clear, easy to use, and provided explicit and unambiguous guidance to assessment.

Phase Three produced a refined rating scale suitable for piloting. The rating guide was also amended to correspond with the refinements.

8. Phase Four, trialling the scale

Twenty qualified ESL teachers, experienced as National Senior Certificate examiners, were involved in piloting the refined scale in Phase Four. They were a convenience sample, but were representative of the population of examiners in that they came from a various L1 backgrounds and schools, which included examiners from historically advantaged as well as disadvantaged schools.

Piloting occurred during a two-day workshop that followed a two-step process involving four on-site scoring iterations. During the first three iterations raters were trained and familiarised with the scale, while the fourth iteration served as final test for inter- and intra-rater reliability in applying the scale. Seven benchmarked essays were randomly selected from each of the seven performance levels to illustrate typical performances across the range of the scale for an initial training session. The first iteration then comprised a blind scoring exercise of four essays randomly selected from the sixty-four benchmarked ones, followed by another training session, during which raters were asked to motivate why they had assigned a particular score to a certain feature. Any differences were discussed. In this way any misconceptions and misconceptions about the scale and its application became evident and could be addressed and clarified through discussion, explanation, and reference to the rating guide and exemplar benchmarked scripts. Training and standardisation continued in this manner after the second and third iterations, during which four and seven additional randomly selected essays were rated respectively.

This training procedure revealed raters' inconsistent interpretation of four micro-categories in particular, viz. number one, insight into and understanding of the topic, number two, originality, number three, mature ideas (Content criterion), and number seven, paragraphing (Structure and development criterion). The distinctions between these items were clarified by emphasising the main focal aspect of each feature as specified in the rating guide. The main source of confusion in the first two of these items appeared to be different interpretations of the role of content in each of these, with some raters equating mature ideas with insight into the topic. Essays would then be penalised for both features, even though some essays, for example, showed insight into the topic (i.e. clear understanding of the topic and providing relevant information), but without expressing mature ideas on it. Furthermore, some raters seemed automatically to equate criterion two (originality) with criterion seven (paragraphing), i.e. if a penalty or credit was assigned to one of these, the other would automatically be penalised or credited accordingly. After examining the scripts, the panel agreed that the two criteria were in fact distinct and that such an equation was not valid.

Two micro-criteria related to paragraphing (numbers three and seven) proved the most problematic, with the most severe discrepancies in raters' scoring. In this case the degree of focus on content had to be clarified. Whereas criterion three explicitly focussed on whether ideas were organised in a logical order, number seven addressed content in the sense of the degree to which the content of each paragraph supported the surface structuring of paragraphs. Finally, it was agreed that 'effective paragraphing' entailed clear organisation of ideas within paragraphs (criterion three), but also clear division of ideas into visible paragraphs (criterion seven). So, if ideas were presented in a logical order, without being clearly organised into paragraphs, criterion seven would be penalised, but not number three.

During the fourth and final iteration, the raters each scored fifteen randomly selected essays individually, without any discussion.

Scores for each of the four iterations were calibrated and analysed statistically. Reliability estimates were calculated for each by means of STATISTICA (StatSoft, Inc., 2008). To investigate inter-rater consistency, the following calculations were done: average inter-rater correlations (Pearson's correlation) with Cronbach's alpha coefficient based on it, Kendall's concordance coefficient, based on Spearman's rank correlation coefficient, and Cronbach's alpha coefficient. Intra-class correlation coefficients were calculated to measure the intra-rater reliability for individual raters (SAS, 2005). To measure the reliability of the panel as a whole, generalisability coefficients were calculated, based on the intra-class coefficient. These also indicated the degree to which the scale could be generalised to other situations, i.e. may be applied consistently by the larger population of examiners and implemented for the purpose of assessing the writing paper. Rasch analyses were conducted for each of the four iterations to determine the reliability of the rating procedure when applied by a single rater and by the group (cf. Stemler 2004; Shaw, 2004).

Scores for each of the iterations were recorded and labelled Batches One to Four. The data generated at each iteration were used to investigate the following:

- the effects of training on inter-rater consistency across the four iterations when participants apply the draft scale;
- inter- and intra-rater consistency concerning the fifteen features and the five criteria;
- relevance of the fifteen criteria, and
- the degree to which individual criteria represent distinct aspects of writing.

Reliability estimate calculations require complete data, i.e. scores reported by each rater for each feature on every script scored. For the purpose of reliability estimate calculations, therefore, only those cases where all raters provided complete scores were used. The number of observations, in terms of the number of essays and the number of raters used for the particular calculation, are indicated in Tables 1 and 2. Table 1 summarises the results for the average inter-rater correlation, Kendall's concordance coefficient and Cronbach's alpha based on it.

Table 1: Results for reliabilities as calculated for each iteration

	Batch 1	Batch 2	Batch 3	Batch 4a	Batch 4b
Number of essays	**4**	**4**	**7**	**12**	**15**
Number of raters	**16**	**17**	**16**	**19**	**17**
Average inter-rater correlation	**0.90**	**0.82**	**0.68**	**0.82**	**0.83**
Kendall's concordance	**0.64**	**0.63**	**0.51**	**0.81**	**0.90**
Cronbach's alpha	**0.97**	**0.96**	**0.95**	**0.98**	**0.99**

Data for Batch 4 were complete for all raters on twelve of the fifteen essays (Batch 4a). Two raters, however, proved inconsistent in their scoring, assigning very low scores where the majority of the panel assigned higher scores. They were removed from the data set. Calculations were then repeated for the remaining data set (Batch 4b) containing complete data for all fifteen scripts. They are reported alongside the original results for comparison.

Average inter-rater correlation coefficients reported in Table 1 demonstrate an initial decrease in inter-rater agreement, followed by a steady increase in the final phases, reaching a high level of inter-rater agreement in the final iteration. This may be a result

of the selection of scripts scored in the second and third iteration. Scripts in Batch 2 more distinctly illustrated performances at different levels, making it perhaps easier to distinguish accurately between them than between the scripts in Batch 3, which mostly contained adjacent performance levels (Levels 3 and 4). Closer examination of the scores assigned by raters for Batch 3 scripts revealed that raters generally differed within one adjacent level of each other on individual features. The increasing tendency in later iterations may indicate that training helped to clarify certain features and standardise interpretation and application of the scale resulting in more consistent scores.

Table 2 reports intra-class correlations, which indicate the degree to which each essay was awarded similar scores by different raters. It also reports generalisability estimates, which indicate the degree to which raters' performances may be interpreted as representative of raters in general. In other words, it provides an answer to the question "Can the results be accepted as indicative of performance of the large population of raters using the proposed scale?"

Table 2: Results for inter-class correlation and generalisability coefficient as calculated for each iteration

	Batch 1	Batch 2	Batch 3	Batch 4a	Batch 4b
Intra-class coefficient for individual raters	**0.37**	**0.58**	**0.30**	**0.74**	**0.82**
Generalisability for the sum of all raters	**0.90**	**0.97**	**0.91**	**0.98**	**0.99**
Cronbach's alpha	**0.95**	**0.97**	**0.92**	**0.98**	**0.99**

Table 2 shows a general increasing tendency in the intra-class coefficients for individual raters, demonstrating that they became more standardised, i.e. more consistent in scoring as training progressed. For the final scoring in Batch 4b, the high value of 0.82 indicates that raters were in close agreement in the scores they awarded. High generalisability coefficients for the sum of all raters on all four batches indicate that the performance of the group of markers can be accepted as indicative of typical performances of examiners in general. There is also a general increase in Cronbach alpha values that resulted in a high alpha value of 0.99 for Batch 4b.

All data sets were then calibrated individually with FACETS Rasch (Linacre, 2006). The results were analysed and compared for the four iterations. In comparing the results, the success of the training procedure became evident, with inter- and intra-rater consistency increasing as well as improved distribution of scale features. This resulted from more accurate interpretation and consistent application of the scale.

After the final calibration, a high reliability index of 0.90 was again reported for raters. Fit statistics identified problematic scores on three different essays for two raters respectively owing to severe variance reported for these two. These raters may benefit from further training, but considering of the aim of this phase, the mis- and overfitting values reported here were not reason for serious concern. Only one slightly misfitting value for micro-criteria was reported for number 15 (Length) with infit mean-square of 1.55 and outfit of 1.64. A very high reliability index of 0.98 for the micro-criteria confirmed that the features were clearly distinguished by raters.

After the fourth and final iteration, the panel of raters reflected on the activities in Phase Four in a discussion. One point that emerged unanimously was that a zero mark option had to be included in the seven-point scale, as well as in the Length criterion, as without it no performance could be assigned a score lower than 15 out of 100 (15 x Level One). This adjustment was made to the scale. Appendix B contains the final version of the scale. The final version of the rating guide is in Appendix C.

Qualitative feedback from the raters on the scale was also obtained from a questionnaire adapted from Shaw and Falvey (2008) after the final iteration. The panel expressed some concern about the number of criteria and the time it might take to score them. Despite these concerns, they unanimously indicated a preference for a detailed scale which would result in accurate scores, as opposed to a fast scoring procedure rendering inconsistent results. The raters stated that they found the scale a clear and simple tool that facilitated objective and consistent scoring, and an improvement on the current scale. They felt that the proposed scale provided a means for a systematic, structured assessment of essays, which could make assessment easier, more precise and faster once raters have been trained in its use.

The MFRM report supported the qualitative results. There was an improvement in rater consistency across the four iterations during training. This suggests that, although the scale may seem complex at first, it may be useful for scoring essays by both more and less-experienced raters. We believe that less-experienced raters in particular may potentially benefit from this type of instrument than from a more holistic grid.

9. Conclusion

This article contains a step-by-step description of the process followed in developing and validating a rating scale for the assessment of writing. Our approach was that any scale must be based on samples of learner writing, and that scale validation involves both a priori and posteriori procedures. We adopted an argument-based validation process instead of an accumulation-of-evidence approach, which is problematic because it is difficult to decide what kind of and how much evidence is necessary (cf. Chapelle et al., 2008:320). Phases One to Four illustrate how one aspect of an evaluation inference prompts empirical development of a rating scale, and contributes to the backing for the assumption and warrant of the inference. The development of the scale amounts to an

argument-within-an-argument, as it is situated within the overall validity argument for the assessment of writing in Grade 12. It is part of the evaluation inference, which in turn forms part of the overall validity argument. The assessment of writing in the Grade 12 examination needs to be investigated as a whole to arrive at a complete validity argument, and the generalization, explanation, extrapolation and utilization inferences need to be investigated to complete a full validity argument for a writing test (for a discussion of each of these, see Chapelle et al., 2010).

Further research on the wider implementation of the scale is, of course, necessary, to provide further backing for its validity. Based on the results reported here, we are confident that the proposed rating scale meets validity requirements and would be an appropriate instrument for the assessment of essay writing in the NSC examination in South Africa.

10. Acknowledgement

We acknowledge financial support from the South African National Research Foundation (NRF) for this project.

References

Alderson, J.C. 1991. Bands and scores. In: Alderson, J.C. and North, B. (Eds.) 1991. *Language testing in the 1990s*. London: Macmillan.

Alderson, J.C., Clapham, C. & Wall, D. 1995. *Language test construction and evaluation. Cambridge*: Cambridge University Press.

Bachman, L.F. 1990. *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L.F. & Palmer, A.S. 2010. *Language assessment in the real world: designing language assessments and justifying their use*. Oxford: Oxford University Press.

Bond, T.G. & Fox, C.M. 2001. *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ.: Lawrence Erlbaum.

Bond, T.G. & Fox, C.M. 2007. *Applying the Rasch model: Fundamental measurement in the human sciences*. 2nd ed. Mahwah, NJ.: Lawrence Erlbaum.

Chapelle, C.A. 1999. Validity in language assessment. *Annual Review of Applied Linguistic,* 19: 254-272.

Chapelle, C.A. 2012. Validity argument for language assessment: the framework is simple... *Language Testing* 29(1): 19-27.

Chapelle, C.A., Enright, M.K. & Jamieson, J.M. 2008. *Building a validity argument for the Test of English as a Foreign Language.* New York: Routledge.

Chapelle, C.A., Enright, M.K. & Jamieson, J.M. 2010. Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice* 29(1): 3-13.

Coniam, D. 2010. Validating onscreen marking in Hong Kong. *Asia Pacific Education Review* 11(3): 423-431. http://www.springerlink.com/content/l6663m43jn501317 Date of access: January 2011.

Davies, A. & Elder, C. 2005. Validity and validation in language testing. In: Hinkel E (Ed.) 2005. *Handbook of research in second language teaching and learnin*g. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Department of Education. 2005. National Curriculum Statement (Grades 10-12). Pretoria: Government Printer.

Department of Education 2008a. Language Programme Guidelines. Pretoria: Government Printer.

Department of Education. 2008b. Subject Assessment Guidelines. Pretoria: Government Printer.

Douglas, D. 2001. Language for specific purposes assessment criteria: Where do they come from? *Language Testing* 18(2): 171-185.

Eckes, T. 2005. Examining rater effects in TestDaF writing and speaking performance assessment: a many-facet Rasch analysis. *Language Assessment Quarterly* 2(3): 197-221.

Engelhard, G. 1992. The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education* 5(3): 171-191.

Fulcher, G. 1987. Test of oral performance: the need for data-based criteria. *English Language Teaching Journal* 41(4): 287-291.

Fulcher, G. 1993. *The construction and validation of rating scales for oral tests in English as a Foreign Language*. Unpublished PhD thesis. Lancaster: University of Lancaster.

Fulcher, G. 2003. *Testing second language speaking.* Harlow: Pearson.

Hattingh, K. 2009. *The validation of a rating scale for the assessment of essays in ESL*. Unpublished PhD thesis. Potchefstroom: North-West University.

Haney, W., Russell, M., Gulek, C. & Fierros, E. 1998. Drawing on education: Using student drawings to promote middle school improvement. *Schools in the Middle* 7(3): 38- 43.

IELTS. 2007. IELTS Handbook. University of Cambridge: ESOL examinations. http://www.cambridgeesol.org/assets/pdf/resources/IELTS_Handbook.pdf. Date of access: March 2008.

Jacobs, H, Zingraf, S.A., Wormuth, D.R., Hartfield, V.F. & Hughey, J.B. 1981. *Testing ESL essays: a practical approach.* Rowley, MA: Newbury House.

Kane, M.T. 2001. Current concerns in validity theory. *Journal of Educational Measurement* 38(4): 319-342

Kane, M. 2004. Certification testing as an illustration of argument-based validity. *Measurement* 2(3): 135-170.

Kane, M.T. 2006. Validation. In: Brennan R (Ed.) 2006. *Educational Measurement*, 4th ed. Westport, CT: Greenwood.

Kane, M.J. 2012. Validating score interpretations and uses. *Language Testing* 29(1): 3-17.

Kane, M., Crooks, T. & Cohen, A. 1999. Validating measures of performance. *Educational Measurement: Issues and Practice* 18(2): 5-17.

Knoch, U. 2009. Diagnostic writing assessment: The development and validation of a rating scale. Frankfurt: Peter Lang.

Language Programme Guidelines. See South Africa, Department of Education. 2008a.

Linacre, J.M. 2002. What do infit and outfit, mean-square and standardization mean? *Rash Measurement Transactions* 16(2): 878.

Linacre, J.M. 2006. *FACETS Rasch Measurement Computer Program Facets for Windows version 3.61.0.* Chicago: Winsteps.com.

McNamara, T. 1996. *Measuring second language performance.* New York: Longman.

McNamara, T. 2000. *Language testing.* Oxford: Oxford University Press.

Messick, S. 1989. Validity. In: Linn, R. (Ed.) 1989. *Educational Measurement, 3rd ed.* New York: Macmillan.

Myford, C.M. & Wolfe, E.W. 2003. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement* 4(4): 386-422.

Myford, C.M. & Wolfe, E.W. 2004. Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement* 5(2): 189-227.

National Curriculum Statement. See South Africa. Department of Education. 2005.

North, B. 2003. Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. TOEFL Monograph 24. Princeton: Educational Testing Service.

North, B. & Schneider, G. 1998. Scaling descriptors for language proficiency scales. *Language Testing* 15(2): 217-263.

Park, T. 2005. An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics* 4(1): 1-21.

SAS. 2005. SAS Institute Inc., SAS Online Doc®, Version 9.1. Cary, NC.

Schaefer, E. 2008. Rater bias patterns in an EFL writing assessment. *Language Testing* 25(4): 465-493.

Shaw S. 2004. IELTS writing: revising assessment criteria and scale (Phase 3). *Research Notes, University of Cambridge ESOL Examinations* 16: 3-6. Cambridge: Cambridge ESOL Examinations.

Shaw, S.D. & Falvey, P. 2008. The IELTS writing assessment revision project: Towards a revised rating scale. Research Report 1: January. Cambridge: University of Cambridge ESOL Examinations.

Shaw, S.D. & Weir, C. 2007. *Examining writing: research and practice in assessing second language writing. Studies in Language Testing 26*. Cambridge: Cambridge University Press.

South Africa, Department of Education. 2005. National Curriculum Statement (Grades 10-12). Pretoria: Department of Education.

South Africa, Department of Education. 2008a. Language Programme Guidelines. Pretoria: Department of Education.

South Africa, Department of Education. 2008b. Subject Assessment Guidelines. Pretoria: Department of Education.

StatSoft, Inc. 2008. STATISTICA (data analysis software system), version 8.0. Available at www.statsoft.com. Date of access: February 2009.

Stemler, S.E. 2001. An overview of Content Analysis. *Practical Assessment, Research & Evaluation* 7(17). http://PAREonline.net/getvn.asp?v=7&n=17. Date of access: January 2009.

Stemler, S.E. 2004. A comparison of consensus, consistency, and measuring approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation* 9(4). http://PAREonline.net/getvn.asp?v=9&n=4. Date of access: January 2009.

Subject Assessment Guidelines. See South Africa, Department of Education. 2008b.

Toulmin, S. 2003. *The uses of argument.* 2nd ed. Cambridge: Cambridge University Press.

Turner, C.E. 2000. Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review* 56(4): 55 – 584.

Upshur, J.A. & Turner, C.E. 1995. Constructing rating scales for second language tests. *ELT Journal 49*(1): 3-12.

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

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

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

Weir C.J. 2005. *Language testing and validation: An evidence-based approac*h. Oxford: Palgrave Macmillan.

Wright B.D. & Linacre J.M. 1994. *Reasonable mean-square fit values. Rasch Measurement Transactions* 8(3): 370.

Appendix A: Current rating scale

		Outstanding	Meritorious	Substantial	Adequate	Moderate	Elementary	Not achieved
ENGLISH FIRST ADDITIONAL LANGUAGE RUBRIC NSC SECTION A: ESSAY 50 MARKS	LANGUAGE	- Language, punctuation effectively used. Uses figurative language appropriately. - Choice of words highly appropriate. - Sentences, paragraphs coherently constructed. - Style, tone, register highly suited to topic. - Text virtually error-free following proof-reading, editing. - Length in accordance with requirements of topic.	- Language, punctuation correct, and able to include figurative language correctly. - Choice of words varied and Correctly used. - Sentences, paragraphs logical, varied. - Style, tone, register appropriately suited to topic. - Text largely error-free following proof-reading, editing. - Length correct.	- Language and punctuation mostly correct. - Choice of words suited to text. - Sentences, paragraphs Well-constructed. - Style, tone, register suited to topic in most of the essay. - Text by and large error-free following proof-reading, editing. - Length correct.	- Language simplistic, punctuation adequate. - Choice of words adequate. - Sentences, paragraphing might be faulty in places but essay still makes sense. - Style, tone, register generally consistent with topic requirements. - Text still contains errors following proof-reading, editing. -Length correct.	- Language ordinary and punctuation often inaccurately used. - Choice of words basic. - Sentences, paragraphs, faulty but ideas can be understood. - Style, tone, register lacking in coherence. - Text contains several errors following proof-reading, editing. - Length – too long / short.	- Language and punctuation flawed. - Choice of words limited. - Sentences, paragraphs constructed at an elementary level. - Style, tone, register inappropriate. - Text error-ridden despite proof-reading, editing. - Length – too long / short	- Language and punctuation seriously flawed. - Choice of words inappropriate. - Sentences, paragraphs muddled, inconsistent. - Style, tone, register flawed in all aspects. - Text error-ridden and confused following proof-reading, editing. - Length – far too long / short

		Outstanding	Meritorious	Substantial	Adequate	Moderate	Elementary	Not achieved
CONTENT		Code 7: 80 – 100%	Code 6: 70 – 79%	Code 5: 60 – 69%	Code 4: 50 – 59%	Code 3: 40 – 49%	Code 2: 30 – 39%	Code 1: 00 – 29%
Outstanding - Content shows impressive insight into topic. - Ideas: thought-provoking, mature. - Coherent development of topic. Vivid detail. - Critical awareness of impact of language. - Evidence of planning and/or drafting has produced virtually flawless, presentable essay.	Code 7 80-100%	40 - 50	38 – 42	35 – 39				

		Outstanding	Meritorious	Substantial	Adequate	Moderate	Elementary	Not achieved
Meritorious - Content shows thorough interpretation of topic. - Ideas: imaginative, interesting. - Logical development of details. Coherent. - Critical awareness of impact of language. - Evidence of planning and/or drafting has produced a well-crafted, presentable essay.	Code 6 70-79%	38 – 42	35 – 39	33 – 37	30 – 34			
Substantial - Content shows a sound interpretation of topic. - Ideas: interesting, convincing. - Several relevant details developed. - Critical awareness of language evident. - Evidence of planning and/or drafting has produced a presentable and very good essay.	Code 5 60-69%	35 – 39	33 – 37	30 – 34	28 – 32	25 - 29		

		Outstanding	Meritorious	Substantial	Adequate	Moderate	Elementary	Not achieved
Adequate - Content: an adequate interpretation of topic. - Ideas: ordinary, lacking depth. - Some points, necessary details developed. - Some awareness of impact of language. - Evidence of planning and/or drafting has produced a satisfactorily presented essay.	Code 4 50-59%		30 – 34	28 – 32	25 – 29	23 – 27	20 – 24	
Moderate - Content: ordinary. Gaps in coherence. - Ideas: mostly relevant. Repetitive. - Some necessary points evident. - Limited critical language awareness. - Evidence of planning and/or drafting that has produced a moderately presentable and coherent essay.	Code 3 40-49%			25 – 29	23 – 27	20 – 24	18 – 22	w15 – 19

		Outstanding	Meritorious	Substantial	Adequate	Moderate	Elementary	Not achieved
Elementary - Content not always clear, lacks coherence. - Ideas: few ideas, often repetitive, - Sometimes off topic. General line of thought difficult to follow. - Inadequate evidence of planning/drafting. Essay not well presented.	Code 2 30-39%				20 – 24	18 – 22	15 – 19	03 – 17
Not Achieved - Content irrelevant. No coherence. - Ideas: repetitive, off topic. - Non-existent planning/drafting. Poorly presented essay.	Code 1 00-29%					15 – 19	03 – 17	00 – 14

Appendix B: Proposed New Rating Scale

RATING SCALE: ESSAY WRITING									
	Poor		Adequate			Very good			
A. CONTENT									
No insight into and understanding of topic.	0	1	2	3	4	5	6	7	Outstanding insight into and comprehensive understanding of topic.
Hardly any originality and/or little interest/ mundane.	0	1	2	3	4	5	6	7	Highly original/ Fresh perspective/ original/ engaging creativity.
Irrelevant and immature ideas.	0	1	2	3	4	5	6	7	Mature and thought provoking ideas.
Does not follow the conventions of essay type.	0	1	2	3	4	5	6	7	Ideally follows conventions of essay type.
Incoherent flow of ideas.	0	1	2	3	4	5	6	7	Highly coherent flow of ideas.
STRUCTURE AND DEVELOPMENT									
No division into introduction, body, conclusion.	0	1	2	3	4	5	6	7	Effective division into introduction, body and conclusion.
No paragraphing.	0	1	2	3	4	5	6	7	Effective paragraphing.
GRAMMAR									
Incorrect syntax.	0	1	2	3	4	5	6	7	Correct syntax.
Incorrect tense & concord.	0	1	2	3	4	5	6	7	Correct tense & concord.
No variety in range of sentence types.	0	1	2	3	4	5	6	7	Wide variety in range of sentence type.
Multiple errors in spelling & punctuation.	0	1	2	3	4	5	6	7	Error-free spelling & punctuation.
VOCABULARY									
Limited range.	0	1	2	3	4	5	6	7	Extended range.
Inappropriate style, diction & register.	0	1	2	3	4	5	6	7	Highly appropriate style, diction & register.

RATING SCALE: ESSAY WRITING									
	Poor		Adequate			Very good			
Ineffective use of linking devices (words & phrases).	0	1	2	3	4	5	6	7	Sophisticated use of linking devices (words & phrases).
LENGTH									
Deviates from requirement.	0		1			2			Adheres to requirement.
TOTAL	100								

Appendix C: Rating Guide for Proposed New Scale

RATING GUIDE: ESSAY WRITING RATING SCALE		
	Criteria & Features	INSTRUCTIONS AND EXPLANANTIONS
	A. CONTENT	This category concerns the ideas presented in the essay in terms of relevance and topicality; novelty, progression and appropriateness.
1	Insight into and understanding of the topic	Assess whether candidate has addressed, developed and sustained the topic.
2	Originality and/or interest	Assess the e xtent to which the essay engages the reader. Give credit for any response that provides fresh/creative perspective on the topic.
3	Relevance and maturity of ideas	The essay must be clearly relevant to the topic. Ideas should be thought through and contribute to the main topic.
4	Appropriateness of structure to text type	The main sections of the essay must follow the conventions of the essay type (argumentative, narrative, descriptive, comparison and contrast, cause and effect).
5	Flow of ideas through the essay	The essay must show natural/ logical progression of ideas/ events/ facts from the introduction to the conclusion and between paragraphs.
	B. STRUCTURE AND DEVELOPMENT	This category refers to the way information is organised in the essay in accordance to the essay type (argumentative, narrative, descriptive, comparison and contrast, cause and effect).
6	Introduction, body and conclusion	The essay must contain a clear introduction, body and conclusion.
7	Paragraphing	The essay must be divided into paragraphs. Each paragraph must have a main idea (usually a topic sentence). The main idea should be developed further by the supporting sentences in the paragraph.

RATING GUIDE: ESSAY WRITING RATING SCALE		
	Criteria & Features	**INSTRUCTIONS AND EXPLANANTIONS**
	A. CONTENT	**This category concerns the ideas presented in the essay in terms of relevance and topicality; novelty, progression and appropriateness.**
	C. GRAMMAR	**This section deals with the accurate use of grammatical structures.**
8	**Syntax**	**As a rule, sentences must be complete (subject & main verb), and contain correct word order. Exceptions used for creative effect should not be penalised if appropriate.**
9	**Use of tense and concord**	**Tense and concord must be used correctly and appropriately.**
10	**Range of sentence types**	**The essay must demonstrate a variety of sentence types and sentences of different lengths and structures accurately and effectively.**
11	**Spelling and/or capitalisation and punctuation**	**Spelling must be accurate (this includes the use of the apostrophe) Capital letters must be used appropriately. If the entire essay is written in capital letters, award a maximum of 3 for category C11. Punctuation (e.g. full stops, commas, colons, dashes and inverted commas) must be used appropriately and correctly.**
	D. VOCABULARY	**This section assesses the extent, accuracy and appropriateness of a candidate's vocabulary.**
12	**Range of vocabulary**	**Candidates have to demonstrate that they have a sufficient extent of vocabulary to express their ideas. Credit must be given for sophistication in words and expressions.**
13	**Appropriateness of vocabulary**	**Words must be used correctly and appropriately. Assess the candidate's ability to use style appropriately, such as formal and informal, narrative, descriptive and argumentative.**
14	**Use of linking words and phrases**	**The candidate demonstrates the ability to use conjunctions, pronouns, adverbs and other devices to link parts of sentences, sentences and paragraphs.**
	E. LENGTH	**The candidate must adhere to the length limitation as specified on the examination question paper.**