Elsevier Editorial System(tm) for Assessing Writing
Manuscript Draft

Manuscript Number: ASW-D-15-00012R2

Title: Developing rubrics to assess the reading-into-writing skills: a case study

Article Type: SI: RUBRICS

Keywords: Reading-into-writing; Writing Assessment; Scoring; Integrated tasks; L2 Writing; CEFR

Corresponding Author: Dr. Sathena Hiu Chong Chan, Ph.D

Corresponding Author's Institution: University of Bedfordshire

First Author: Sathena Hiu Chong Chan, Ph.D

Order of Authors: Sathena Hiu Chong Chan, Ph.D; Chihiro Inoue, Ph.D; Lynda Taylor, Ph.D

Abstract: The integrated assessment of language skills, particularly reading-into-writing, is experiencing a renaissance. The use of rating rubrics, with verbal descriptors that describe quality of L2 writing performance, in large scale assessment is well-established. However, less attention has been directed towards the development of reading-into-writing rubrics. The task of identifying and evaluating the contribution of reading ability to the writing process and product so that it can be reflected in a set of rating criteria is not straightforward. This paper reports on a recent project to define the construct of reading-into-writing ability for designing a suite of integrated tasks at four proficiency levels, ranging from CEFR A2 to C1. The authors discuss how the processes of theoretical construct definition, together with empirical analyses of test taker performance, were used to underpin the development of rating rubrics for the reading-into-writing tests. Methodologies utilised in the project included questionnaire, expert panel judgement, group interview, automated textual analysis and analysis of rater reliability. Based on the results of three pilot studies, the effectiveness of the rating scales is discussed. The findings can inform decisions about how best to account for both the reading and writing dimensions of test taker performance in the rubrics descriptors.

Detailed Response to Reviewers

We have removed the yellow highlights and double spaced the paper.

***Title Page (WITH Author Details)**

Developing rubrics to assess the reading-into-writing skills: a case study

Sathena Chan*, Chihiro Inoue, Lynda Taylor

Centre for Research in English Language Learning and Assessment, University of Bedfordshire, Hitchin Road, Luton, LU2 8LE, United Kingdom

*Corresponding author, Tel.: +44 01582 489795
E-mail address: sathena.chan@beds.ac.uk

***Brief author biography**

Sathena Chan is a Lecturer in Language Assessment at CRELLA, University of Bedfordshire. Her research interests include integrated reading-into-writing assessment, test development and validation, and cognitive processing of language use. She has worked on a range of test development and validation projects for examination boards and educational organisations in the UK and worldwide.

Chihiro Inoue is a Lecturer in Language Assessment at CRELLA, University of Bedfordshire. Her main research interests lie in the test task design, rating scale development and the criterial features of learner language. She has worked on a number of test development and validation projects in English and Japanese languages in the UK, USA and Japan.

Lynda Taylor is a Senior Lecturer at CRELLA, University of Bedfordshire. She holds an MPhil and PhD in language testing and assessment, both from the University of Cambridge, UK. Over 30 years she has accumulated extensive knowledge and experience of the theoretical and practical issues in language teaching, learning and assessment.

Highlights

- The project developed and validated a suite of reading-into-writing rubrics.
- The project adopted an empirical mixed-method approach to developing rubrics.
- The project provided validity evidence that the rubrics have met key quality standards.

Abstract

The integrated assessment of language skills, particularly reading-into-writing, is experiencing a renaissance. The use of rating rubrics, with verbal descriptors that describe quality of L2 writing performance, in large scale assessment is well-established. However, less attention has been directed towards the development of reading-into-writing rubrics. The task of identifying and evaluating the contribution of reading ability to the writing process and product so that it can be reflected in a set of rating criteria is not straightforward. This paper reports on a recent project to define the construct of reading-into-writing ability for designing a suite of integrated tasks at four proficiency levels, ranging from CEFR A2 to C1. The authors discuss how the processes of theoretical construct definition, together with empirical analyses of test taker performance, were used to underpin the development of rating rubrics for the reading-into-writing tests. Methodologies utilised in the project included questionnaire, expert panel judgement, group interview, automated textual analysis and analysis of rater reliability. Based on the results of three pilot studies, the effectiveness of the rating scales is discussed. The findings can inform decisions about how best to account for both the reading and writing dimensions of test taker performance in the rubrics descriptors.

1. Background to the study

Much real-world writing is composed in response to a text (or texts) requiring high-level reading skills to integrate the input materials into the written response (AUTHORS, DATE; Gebril, 2009; Gebril and Plakans, 2009; Plakans and Gebril, 2012; Weigle, 2004). For example, assignments at schools and universities often require reading multiple texts (e.g. books and articles), gathering information, developing thoughts, and then writing to produce an organised response which incorporates selected information from the sources. With an increasing number of international students who wish to study in English-medium courses and programmes, there has been a growing interest in the 'integrated' assessment of language skills in recent years, which enables stakeholders to infer how well a test taker may be able to handle this type of writing in real life. In this paper, such writing tasks are called "reading-into-writing[1]" (Weigle, 2004), and they refer to single tasks that require students to write a continuous text by drawing upon single or multiple reading materials which can be verbal, non-verbal or both. This integrated reading-into-writing task type has the potential to satisfy the need for greater validity in the assessment of test takers' writing ability as such a task type represents more closely how people write in real life than independent writing tasks (e.g. Cumming, Grant, Mulcahy-Ernt, & Powers, 2004; Cumming, Kantor, & Power, 2001; Weigle, 2004).

Accordingly, some testing organisations such as Educational Testing Service and Trinity College London (hereafter Trinity) have developed and used reading-into-writing tasks in their tests. This enables closer investigation of the **reading-into-writing construct**, including consideration of how appropriate **rubrics** can be developed. This article reports a recent project

[1] Although some researchers use the terminology 'reading-to-write', the term 'reading-into-writing' is usually used by large-scale testing providers as a category to refer to this task type. As this study is primarily concerned with language testing, the terminology of 'reading-into-writing' rather than 'reading-to-write' is used throughout.

to develop and validate rubrics for assessing the skills of reading-into-writing within Trinity's suite of Integrated Skills of English (ISE) examinations, which involve integrated tasks at four proficiency levels - ISE Foundation, ISE I, ISE II and ISE III. The four levels of ISE are targeted to align[2] with the levels of the Common European Framework of Reference (CEFR) for Languages from A2 (Basic User – Waystage), B1 (Independent User – Threshold), B2 (Independent User - Vantage) to C1 (Proficient User – Effective) (for more details, see Council of Europe, 2001). ISE has been designed to assess proficiency levels of test takers who are either in or entering into an educational context, and the "intended candidate is a young person or adult, typically at secondary school or college who is using English as a second or foreign language as part of their studies in order to develop their skills and improve their knowledge of a range of subject areas" (Trinity College London, 2015a: p.7). According to Trinity (2015a), ISE qualifications can be used as a proof of English proficiency for entering university or employment, enrolling into higher level of English study or further education, and/or for UK visa application purposes.

ISE consists of two modules, namely, *Reading and Writing* and *Speaking and Listening*. The work reported in this article is on the former, and is part of a larger ISE redevelopment project, the overall aim of which was to revise and update the original ISE suite[3]. Prior to the project, Trinity conducted a needs analysis for the redevelopment of ISE which involved (a) two independent academic reviews of the original examination (AUTHOR, DATE; Green, 2013), (b) market research, (c) focus group interviews with original ISE teachers and raters. Based on the outcomes, the research team decided in collaboration with Trinity's test redevelopment team

[2] The original ISE suite was calibrated to the CEFR during a two-year period between 2005 and 2007. A full account of the calibration process can be found in Papageorgiou (2007), available on the Trinity College London website.

[3] The original ISE suite is available until 31 August 2015. The revised ISE exams are available from 6 April 2015 in the UK and 1 September 2015 outside the UK (Trinity College London, 2015b).

to prioritise the following recommendations for redeveloping the rubrics for the ISE reading-into-writing task (see Table 1).

The aim was to develop a suite of 4 sets of level-specific analytic rubrics for the reading-into-writing and independent writing tasks at ISE F (A2), ISE I (B1), ISE II (B2) and ISE III (C1). This paper will only address the rubrics for the reading-into-writing task. The set of rubrics for the reading-into-writing task[4] at each ISE level was to have four analytic criteria and four bands[5] to indicate *an inadequate performance at the level*, *an adequate performance at the level, a good performance at the level*, or a *strong performance at the level*.

Three key stages were planned for developing and validating the rubrics: defining the theoretical construct, developing the rubrics and validating the rubrics with empirical analyses of test taker performance, rater feedback and rater reliability. Methodologies utilised in the project included questionnaire, expert panel judgement and automated textual analysis. A suite of reading-into-writing rubrics was produced for use in the live test operation of ISE. The findings from this developmental work may help to inform decisions about how best to frame reading-into-writing activity through the task setting and instructions, as well as how to account for both the *reading* and *writing* dimensions of test taker performance in the rubrics descriptors.

2. Literature review

2.1 *General principles of rating scale development*

[4] The wording in the rubrics in this article is from the development phase and may be different from those in the final version published on Trinity's website.

[5] The 4 bands were labelled as Bands 0, 1, 2 and 3 during the development and validation phase of the ISE reading-into-writing rubrics, and thus are referred to as such throughout this paper. During the pre-testing phase (which is beyond the scope of this paper), the 4 bands were renamed as Score 1 (the original band 0), 2 (the original band 1), 3 (the original band 2) and 4 (the original band 3). Score 0 was then used to distinguish scripts which do not need further rating.

Traditionally, the design and development of scoring rubrics (or rating scales) for direct tests of writing and speaking ability depended largely upon a *'a priori'* approach, according to which assessment criteria and rating scale descriptors were developed through connoisseurship, by an individual 'expert' or small group of experts. Fulcher (2003) characterised this as the 'armchair' approach in which experts, such as teachers, applied linguists and language testers, used their own intuitive judgement to isolate key features of writing performance and to hypothesise verbal descriptors of performance quality. This approach has been criticised on the ground that it lacks empirical support and tends to produce decontexualised descriptors (Hudson, 2005; North, 2000), and thus, the 1990s saw increasing calls for a more empirically-based approach to rubric development (Shohamy, 1990; Upshur and Turner, 1995; Fulcher, 1996; McNamara, 1996; North, 2000). Such an approach involves analysing samples of actual language performance, shaped by a specific purpose and context, in order to construct (or reconstruct) the essential assessment criteria and to describe meaningful levels of performance quality. This approach also seeks to take account of how assessment criteria and level descriptors are likely to be interpreted and applied by human raters. The strength of this approach lies in the practicality and authenticity of the resulting rubrics.

More recently, the value of a mixed method approach has been increasingly recognised in order to collect further evidence for rubrics validation from different perspectives (e.g. Cumming et al, 2001; Lim, 2012; Shaw and Weir, 2007). Specifically, this approach combines the intuitive approach with both quantitative and qualitative analyses so that each method contributes in a complementary and cumulative manner to rating scale development and validation. While quantitative methods focus on statistical analyses (e.g. of score data, rater variability) and

interpretation of their findings, qualitative methods can offer valuable insights into how test takers and raters perceive and approach the assessment event (e.g. priorities and processes). Work on the CEFR (Council of Europe, 2001) highlighted some key practical considerations when developing scale descriptors, including positiveness, definiteness and clarity, independence, and brevity, since features such as these can impact directly on the rating process itself, especially the extent to which raters can successfully interpret and apply the rubrics.

The research reported in this paper drew upon a range of intuitive, qualitative and quantitative methods and thus illustrates the current mixed methods paradigm that underpins much of the contemporary research to develop and validate rating rubrics.

2.2 *Rating rubrics for reading-into-writing performance*

In the field of writing assessment research, the scoring validity of integrated reading-into-writing tasks has received much less attention than independent writing-only tasks. Developing specific rubrics for integrated tasks does not seem to be common practice in the industry. Integrated tasks have been used in a number of standardised language tests, such as the General English Proficiency Test (GEPT) and Cambridge English tests such as Cambridge Advanced (CAE) and Cambridge Proficiency (CPE). However, most tests apply a common set of rubrics to their independent and integrated tasks.

A handful of studies have investigated rater processes on integrated reading-into-writing tasks and the difficulties they faced (e.g. Cumming et al., 2001; Green, 1998; Gebril and Plakans, 2014). By using think-aloud protocols, Green (1998) investigated rater processes when rating an integrated reading-into-writing task in CAE. She categorised the processes according to scoring

behaviour, text features, evaluative response, and meta-comments. Raters in her study were most commonly engaged in essay reading and rating processes related to language accuracy, language appropriateness, and task realization/content. However, the specific focus of her work means there was not much discussion of how raters handled the unique features of integrated reading-into-writing performance. In contrast, Cumming et al's study (2001) set out to compare rater processes on the integrated and independent tasks of TOEFL iBT. Based on the think-aloud protocols of seven raters, they found that raters focused more on rhetoric and content when rating integrated tasks, whereas they focused more on language when rating independent tasks. Most recently, Gebril and Plakans (2014) undertook an in-depth analysis of rater processing when rating a reading-into-writing task. Two raters were asked to rate a subset of 21 essays, representing performance at five score levels selected from a larger pool of essays. It should be noted that the design of the study deliberately did not provide the raters with guidance on how they should rate the performance. Based on think-aloud protocols and interview data, the results suggested that raters employed judgement strategies more than interpretation strategies. In more specific terms, they approached the tasks by 'locating source information', 'checking citation mechanics', and 'evaluating quality of source text'. Information of how raters in this study approached the integrated tasks would be helpful for standardising rater processes in rater training.

There is also a need to investigate what difficulties and challenges the raters in this study experienced, so that measures can be taken to reduce these hurdles. Difficulties commonly associated with rating such performance are questions pertaining the quality and style of source text integration, textual borrowing, time effectiveness and task specificity. Cumming et al (2001) noted that raters in their study faced a new set of demands, for instance, to distinguish how well

candidates *transform* source materials in their writing. Difficulties identified by the raters in Gebril and Plakans' (2014) study included (1) distinguishing between language cited from source materials and language produced by writers, (2) assigning a score to essays with a high incidence of quotations, (3) the perception of inappropriate textual practices, and (4) scoring essays that matched the profile of two adjacent levels.

As indicated by the above studies, the need to develop specific rubrics to address the features which are unique to integrated task is evident. This study sets out to investigate the following three research questions:

1. To what extent does analysis of actual test takers' performance help to refine the rubrics*?*
2. What are the raters' perspectives on using the rubrics?
3. To what extent can the rubrics reliably distinguish reading-into-writing performance at different levels?

3. Method

3.1 *Timeline, goals and procedures*

Key stages of the project which are relevant to the discussion of this paper are shown in Table 2:

To develop the suite of analytic rubrics, the research team consulted a range of materials including relevant CEFR descriptors, the original ISE's holistic and generic descriptors, and other established reading-into-writing rating rubrics, e.g. TOEFL iBT holistic integrated writing

scoring rubrics (ETS, 2004). The research team determined the sub-categories under each of the four criteria: *Reading for Writing, Task Fulfilment, Organisation and Structure,* and *Language Control*. (Apart from *Reading for Writing*, the other three criteria are shared with the independent writing task. Therefore, the descriptors for the three criteria needed to be sufficiently generic to be applicable for the two different types of writing.) The analytical descriptors were partially derived from the original ISE performance descriptors. The new descriptors were cross-referenced to the CEFR descriptors at the relevant levels and informed by scrutiny of actual written performances obtained from the Mini-pilot. After the initial draft of the rubrics had been produced, it was piloted with four raters in Pilot 1. A focus group meeting was conducted to collect their feedback. The set of rubrics was revised and then piloted with 12 raters in Pilot 2. A questionnaire was used to collect raters' feedback on the revised rubrics. Rater reliability was calculated by Rasch analysis.

3.2 *Reading-into-writing tasks and test takers*

The reading-into-writing tasks used in this study were the prototype tasks[6] of the new ISE Reading-into-writing task at the four ISE levels. Table 3 provides an overview of the task requirement and target processes for the reading–into-writing task at each level (for a sample of the actual task at ISE I, see Appendix 1). Examples of topics include: place of study (ISE F), means of transport, recent personal experience (ISE I), education, pollution and recycling (ISE II), the school curriculum and scientific developments (ISE III). A fuller list of topics can be found in Trinity College London (2015a).

[6] The details of the prototype tasks provided here may be different from the operationalised ISE Reading and Writing tests. Details of the new ISE Reading and Writing tests are available from Trinity's website.

A total of 642 test takers participated in the three pilots (Mini-pilot, Pilot 1 and Pilot 2). They were recruited from different test centres in India, China, Italy, Spain, UK and Bulgaria to represent ISE's international candidature (see Table 4 for the number at each level). Most of them were secondary school students, and the average age was 14.14 (SD = 5.25). 46.4% of the test takers were male and 53.6% were female. As they had been trained and prepared for the original ISE exams which include a reading-into-writing task in the same format (with less number of source materials[7] than the revised exams), they were reasonably familiar with this task type.

3.3 *Raters and approach to rating*

One experienced EAP lecturer was recruited by the research team to assist with script analysis. The rater and three members of the research team analysed the 40 reading-into-writing scripts collected in the Mini-pilot by (a) ranking the 10 scripts at each ISE level according to two performance bands, i.e. *at or above* and *below* the level, (b) selecting extracts from the script pool to exemplify the 'stronger' and 'weaker' performance, and (c) providing a rationale for the selection. The findings were used to inform the development of the rating descriptors.

The same procedures were repeated to provide an initial analysis of the 186 scripts collected in Pilot 1, but this time the team ranked the scripts at each ISE level into four bands. Features which helped to distinguish the performance at each sub-level were identified to refine the descriptors (see Section 3.4).

[7] The numbers of source texts/materials in the revised ISE exams (i.e. 3 for ISE F, 4 for all other levels) were decided based on the previous test takers' performance, market needs from Trinity's candidature and expert panels' opinions within and outside Trinity.

After that, the scripts were rated in two rounds by a team of four experienced raters recruited by Trinity. They received some initial rater training on the emerging performance descriptors and each rated about a quarter of the scripts at each ISE level. A set of 32 scripts (eight from each level) was held back to be marked in the second round. These 32 scripts were quadruple-marked by the four raters. The four raters then participated in a group interview and their feedback was used to further refine the rating rubrics and descriptors (see Section 3.5).

Following Pilot 2, the scripts were also rated in two rounds. 12 experienced raters were recruited and each rated about one-twelfth of the scripts. All raters filled in a feedback questionnaire (see Section 3.6). Six of the raters then participated in a second round of rating, where all scripts were double rated (see Section 3.7 for the reliability analysis).

3.4 Analysis of reading-into-writing performance (manual and automated)

The scripts collected in the Mini-pilot and Pilot 1 were analysed manually by an independent rater and three members of the research team to a) identify textual features which distinguish the performance at each ISE level and at different bands within each ISE level; and b) to verify the relationship between performance and the rubrics. The scripts were classified into piles based on their scores from the first round marking. The independent rater and the three research team members, functioning as an expert panel, decided which features to include in the rubrics. The expert panel also provided written feedback regarding the distinctive features of the scripts.

In addition, the 186 scripts collected in Pilot 1 were typed up by a research assistant for the purpose of automated textual analysis. 10% of the typed scripts were checked by the research team for accuracy. Automated textual analysis has been regarded as a more systematic and

efficient way to assess textual features than the more traditional expert judgement method, especially when a large number of texts are involved. Automated textual analysis was conducted to give an initial indication of the level of complexity of the reading-into-writing performance at the four ISE levels. Seven indices from Vocabprofile (Cobb, 2003), which is a freely available textual analysis tool, were used. It provides a profile of texts in terms of different vocabulary frequency bands based on the British National Corpus (e.g. the most frequent 1000 words) and different types of vocabulary (e.g. academic words based on Coxhead, 2000). The tool has been widely used to assess the difficulty level of reading texts in the test development and validation projects (e.g. Green, 2012; Wu, 2014).

3.5 Group interview

After the four raters had completed the rating procedure, a focus group interview was conducted to explore their feedback. They were asked to describe how they applied the new rubrics, to identify the strengths and weaknesses of the rubrics used in Pilot 1, to identify any difficulties they experienced and to propose further appropriate revisions for Pilot 2. The interview was recorded and the major findings are discussed in Section 4.2.

3.6 Questionnaire

A feedback questionnaire was administered to the 12 raters who participated in Pilot 2. Five of the questions are relevant to the discussion of this paper and rater responses to these are discussed in Section 4.2.

3.7 Analysis of rater reliability (Rasch analysis)

Score analysis was conducted with the data from Pilot 1 and Pilot 2 in order to examine the effectiveness of the rubrics and rater reliability. As mentioned earlier, the scripts were rated in two rounds: first round by a single rater and then second round by quadruple-marking in Pilot 1 and double-marking in Pilot 2. For each pilot, the analysis included two strands: 1) calculating the average of scores at each ISE level to see the overall trends (using ratings from the first round), and 2) calculating rater reliability and running a Rasch analysis on the scores in order to examine rater consistency and whether rating scales were functioning well (using ratings from the second round). The scores given by multiple raters were analysed using descriptive statistics, traditional rater reliability index (i.e. Spearman-Brown Prophecy Formula (Henning, 1987)) and analysis of rater reliability and rating scale use using Rasch analysis (with FACETS software). The results are reported in Section 4.3.

4. Results

4.1 Using actual test takers' performance to refine the initial rubrics

As described previously, the suite of rubrics consists of four analytic criteria and several sub-categories within each criterion (see Table 6[8]). Each sub-category then consists of descriptors specifying the level of performance expected at each band (i.e. 1: *an inadequate performance at the level*, 2: *an adequate performance at the level*, 3: *a good performance at the level*, or 4: a

[8] The final version of the analytic criteria is presented here for the sake of clarity of this paper.

strong performance at the level). The published version[9] of the rubric at ISE II is provided in Appendix 2. Rubrics at other ISE levels are available on Trinity's ISE website. Whereas the analytic criteria and the sub-categories were the same for all ISE levels, the descriptors at each band were different across the levels. There was a need a) to verify whether all four criteria can be applied at the four ISE levels, and b) to refine the proposed performance descriptors using actual test takers' performance on the reading-into-writing tasks.

The expert panel group provided written feedback on two piles of scripts, i.e. 'stronger' and 'weaker' performance piles (see 3.4 above). The expert panel first compared the written feedback of the two piles and discussed the distinctive features between the stronger and weaker performance according to each criterion. Some examples of the written feedback on the *Reading for Writing* criterion are provided in Table 7 for illustrative purposes (for samples of the actual scripts, see Appendix 3). The refined rubrics were used in the second round of rating (i.e. quadruple-marked by four raters). The ratings from second round were submitted for statistical analyses (see Section 4.3 for the results).

Drawing upon the descriptive feedback on the observed test takers' performance, the research team were able to refine the rubrics in several ways.

Generally speaking, the research team felt that all criteria, i.e. *Reading for Writing*, *Task Fulfilment*, *Organisation and Structure*, and *Language Control* can be applied at all four ISE levels. One change was made to the sub-categories. The research team noted comments regarding test takers' good understanding or misinterpretation of the sources in the scripts but they felt that this was not addressed by any of the four sub-categories under the analytic criterion *Reading for Writing* at that time (i.e. *Selection of relevant content from source texts*, *Ability to identify common themes and links within and across the multiple texts*, *Adaptation of content to*

[9] See footnote 5 above.

suit the purpose for writing and *Use of paraphrasing and/or summarising*). The same issue was discussed in the raters' group interview (for which the results are discussed in Section 4.2). Hence, an additional sub-category *Understanding of reading materials* was added.

The research team found the descriptive comments useful to indicate the features of performance at each ISE level. One of the primary goals of this rubrics redevelopment project was to develop an analytic criterion which can specifically address integrated reading-into-writing abilities that are not assessed on an independent writing-only task. As noted earlier, reading-into-writing analytic rubrics are very rare in the field because most language tests use a common scale for both their independent writing-only and integrated reading-into-writing tasks. It was difficult to develop descriptors for all sub-categories without any reference to the actual test takers' performance.

Another goal was to develop a suite of level-specific analytic rubrics which not only specify performance at each ISE level but also place that performance into one of several possible band levels within each ISE level. To achieve this, the researchers used the written comments to enhance the descriptors at Band 2 (which indicates a good performance at the target level of the examination) and Band 0 (which indicates performance quality falling below the target level of the examination) with reference to the actual test taker's performance. For example, in terms of *Reading for Writing*, Band 2 scripts at ISE I showed evidence of test takers' ability to identify the main conclusions and significant points of the source materials, whereas Band 2 scripts at ISE II offered evidence of test takers' ability to identify common themes across the multiple texts and finer points of details. Another distinct feature regarding the *Reading for Writing* criterion was that Band 2 scripts at ISE F confirmed test takers' ability to retell key ideas, whereas Band 2 scripts at ISE II demonstrated test takers' ability to paraphrase and summarise

factual ideas and opinions. Regarding *Organisation and Structure*, Band 0 scripts at ISE F manifested heavy use of incomplete sentences and lack of any cohesive devices, whereas Band 0 scripts at ISE III were found to lack structure, with limited or inappropriate use of paragraphing, and to lack coherence (i.e. ideas not progressing logically). The research team incorporated these features into the Band 0 and Band 2 descriptors where possible. When the same procedures were repeated, the researchers focused more on the Band 1 and Band 3 descriptors.

In addition to the expert panel's written feedback regarding the distinct features of the reading-into-writing performance, length of scripts in Pilot 1 at each ISE level was analysed to verify the measurement of fluency which is included in the criterion of *Task Fulfilment* under the sub-category of *Adequacy of task coverage.* Number of words is used to reflect the temporal aspect of fluency. Table 8 indicates the mean task response length[10] at each ISE level. The ISE Reading and Writing Exams do not specify time allowance for individual tasks but the task instructions indicate that test takers should spend 40 minutes on this particular reading-into-writing task. The suggestive time is the same for all ISE levels. The figures in Table 8 show a steady increase of number of words provided by test takers at each ISE level. This indicates an increase in test takers' writing fluency across the ISE levels. Other linguistic textual features are addressed by *Language Control.*

To verify the linguistic features and the descriptors of *Language Control*, automated textual analysis was used additionally. Table 9 shows the results of the automated analysis of the 186 scripts collected in Pilot 1, i.e. 45 ISE F, 58 ISE I , 34 ISE II and 49 ISE III scripts. The results showed that the lexical complexity level of the performance increased across the ISE levels,

[10] One of the reviewers suggested that the text length and ratings might be highly correlated. We checked this using non-parametric correlation coefficients, but the correlations between the two were almost non-existent. This may be due to the fact that the instructions state the word limits at each ISE level, and longer responses did not necessarily mean 'better' performances.

although some inconsistency was shown at ISE II. The comparatively fewer scripts at this level may contribute to the inconsistency. (A follow-up analysis was conducted to investigate this issue by visiting the textual complexity of the source texts but is not reported here due to space constraints.) Nevertheless, it was felt that manual analysis by the expert panel appeared to provide more useful information of test takers' performance for refining the rubrics at this initial stage. Apart from actual test takers' performance, raters' feedback, which will be discussed next, was another important source for refining the rubrics.

4.2 *What are the raters' perspectives on using the rubrics?*

Drawing upon the focus group interview and questionnaire data, raters' processing is discussed, followed by their reaction to the new rubrics. Most raters welcomed the three main changes to the overall scoring approach for the ISE tests: (1) from holistic to analytic, (2) from generic to level-specific, and (3) the use of an additional criterion on the integrated task. However, for the purpose of this article, we focused on the issues regarding the initial rubrics raised by the raters.

Rating processes

The raters were asked to describe the processes they used to rate the integrated performance. The major processes which emerged from their discussion include *reading the task instruction*, *reading the rubrics*, *reading the source texts*, *identifying relevant parts in the source texts*, *reading the script*, *assigning a score to each criterion*, *checking the source texts*, *checking the*

rubrics and *reconsidering the assigned scores*. According to the raters, they read the task instruction carefully in order to understand what test takers were required to do. They specifically mentioned the requirements of *using your own words, using the information you read* and some language functions required such as *describe, explain* and *suggest*. Most of them approached the rubrics by reading the sub-categories of each criterion first, followed by Band 2 and then Band 0 descriptors of these sub-categories. They explained that this would help them to understand the different foci of each criterion. They relied on Band 2 descriptors to picture the level of performance required at a particular ISE level, and used Band 0 descriptors to determine what features were below the level. Most of them paid less attention to Band 1 and Band 3 descriptors. All raters read the source texts to identify the necessary content for the task. Some highlighted the relevant parts in the source texts whereas one rater created a simple hierarchy map. The raters tended to read and assess the scripts several times, each time focusing on one criterion, to give the test taker a 'fresh and fair' chance on each criterion. While the raters were only required to assign an overall score for each criterion, two raters tried to assign a score to each sub-category. All raters reported that they checked the source texts again while rating for *Reading for Writing*. They did so especially when they had difficulties in distinguishing between language in the source texts and language produced by the test taker. This problem was also reported in Gebril and Plakans' (2014) study. All raters reported reconsidering scores which they previously assigned to monitor their own consistency. The focus group discussion revealed the general approach and major processes employed by the raters in this study. In the near future, a more in-depth study should be conducted using think-aloud to further investigate rater processes and to explore the several issues raised such as how raters create a representation of the task,

how raters assign the score for each criterion, and whether different sub-categories play a different role across ISE levels.

Difficulty in getting used to the new features

The raters in this study did not seem to have much difficulty getting used to the new analytic rubrics when compared to the original holistic scale. They felt that the analytic descriptors gave them explicit guidance and thus they relied less on their own holistic impression while rating. One rater commented that 'when several experienced raters mark the same script, it is likely that you get a similar score but they may well have very different reasons. Using an analytic scale would give you more confidence that the raters are looking at the same things'.

Nevertheless, they found it much harder to get used to the new criterion - *Reading for Writing.* One rater explained that 'I used to score anything beyond organisation and language under *Task fulfilment*. The new criterion *Reading for writing* addresses different aspects of integration skills which were not mentioned in the old rubrics. It takes time to adjust to the change.' Other raters also reported that they were confused at times between the criteria of *Reading for writing* and *Task Fulfilment*. They thought that the descriptors for these two criteria did not distinguish between each of the bands as well as the other two criteria. They were asked to identify those descriptors which were ambiguous to them and discussed how they interpreted each sub-category of *Reading for writing* and *Task Fulfilment*. A summary of the discussion is presented in Table 10. Such an exercise was useful to clarify and thus minimize the potential overlaps between the two criteria. The discussion outcomes were adapted in rater training materials.

Another concern raised was related to *Language Control*. A list of the target language items at each ISE level is provided in the original ISE syllabus and rubrics. However, recommendations were made to remove the language item list to shift the emphasis from eliciting 'language items' to 'communicative language functions' so as to encourage positive washback in classroom learning. The raters agreed to this rationale and welcomed the potential benefits of the change. Nevertheless, they found it sometimes difficult to use the new rubrics due to the fact that that they no longer had a list of language items to refer to. Additional rater training is clearly essential for raters who are used to working with the original rubrics.

Procedural aspects

The raters identified a few areas for improvement related to procedural aspects of the rubrics. First, they were asked to underline terms or phrases which were not accessible to them, e.g. *transformation* and *synthesis.* These terms were either rephrased or supplemented with additional explanation. Second, they pointed out some inconsistent use of adjectives in the descriptors. Changes were made accordingly. Finally, the raters suggested quantifying some of the descriptors to make the indicated requirements more transparent and hence to make their rating more reliable. Drawing upon the expert panel's written comments on the features of the scripts, additional details were added to the descriptors where deemed necessary. For example, the descriptor *'inadequate and inaccurate selection of relevant content from the source texts*' under the *Reading for Writing* at ISE II means that *'fewer than half of the relevant ideas are selected and most of the selected ideas are irrelevant'*.

Relevance of understanding of source materials

The raters engaged in some debate on whether and how understanding of source materials can be assessed on the integrated task. Some raters suggested excluding this sub-category because comprehension of the source materials cannot be reflected directly, whereas other raters pointed out that some scripts showed clear evidence of misunderstanding of source materials. In addition, as shown in the previous sub-section, the features highlighted by the expert panel include references to test takers' understanding or misinterpretation of the source texts. For this reason, the sub-category of *Understanding of source materials* was kept.

Need to be familiar with the source materials and issues of source use

Raters agreed that there was a need to familiarise themselves with the source texts. Some raters suggested providing a list of the relevant ideas related to the task in the rater training pack. While this might make the remedial support too task-specific and hence less feasible in a standardised setting, it is essential to provide rater training with specific guidelines on the features of the source materials. It is also essential for raters to understand the criterial demand of interaction with the source materials that is expected at each ISE level. The raters then discussed issues related to appropriate/inappropriate source use. The first one was how much lifting was allowed. The raters felt that heavy lifting was not a major issue during in the pilots. This could be because it is stated in the task instructions that students should use their own words. Nevertheless, they expected more occurrences of heavy lifting in the live test. All four raters agreed that scripts with heavy lifting should receive Band 0 (i.e. *an inadequate performance at*

the level) on *Reading for Writing* whereas Band 2 scripts should have almost no lifting but show good evidence of direct quoting, paraphrasing and/or summarising skills. Another interesting issue was test takers' notion of 'use your own words'. The raters felt that some test takers misinterpreted this in a way that they should not repeat any words from the source texts. One rater commented that 'some candidates have obviously understood the texts though [they] don't refer to it directly' whereas 'others were trying too hard to use their own words'. They argued that there is a need to improve students' awareness of appropriate source use. In addition, they argued that raters need more support in order to make better judgement with regards to quality of source texts. They asked for samples of direct quoting, paraphrasing and/or summarising at each ISE level. While some samples were included in the rater training materials, a further validation study on textual features and source use across ISE exams has been scheduled to take place to gather such evidence.

The set of rubrics was refined based on raters' feedback and analysis of rater reliability (see Section 4.3). A questionnaire was then administered to the twelve raters who used the revised rubrics in Pilot 2. The results are presented in Table 11. The results were positive overall. The feedback collected in Pilot 2 seemed to confirm that the revisions made to the rating criteria improved the clarity of the rubrics. A majority of the raters (above 75%) agreed that the descriptors were easy to understand and interpret for all four categories: *Reading for Writing, Task Fulfilment*, *Organisation and Structure*, and *Language Control*. They felt that the descriptors of these categories distinguished well between each of the bands (i.e. Bands 0, 1, 2 and 3) within each ISE level. In addition, most raters agreed that each category was distinct from the other three categories, and the written scripts provided a sufficient quantity/quality of the test

takers' language sample for the rating procedure. The four raters who participated in Pilot 1 agreed that the revised descriptors addressed the major issues they identified in the previous interview. The following section reports findings on rater reliability.

4.3 *To what extent can the rubrics reliably distinguish reading-into-writing performance at different levels? Rater reliability (Rasch)*

Scoring reliability was evaluated using the data gathered from Pilot 1 and Pilot 2. Firstly, inter-rater reliability was calculated using Spearman-Brown Prophecy Formula (Henning, 1987). Table 12 shows the inter-rater reliability of the four raters who participated in Pilot 1 and Table 13 shows those of the six raters who participated in Pilot 2.

The inter-rater reliability was high in all the rating categories across different levels for Pilot 1 (0.80 or above), except for the category of *Organisation and Structure* at ISE II. This is where the raters showed relatively large disagreement on the scores. Drawing upon the raters' group interview and expert panel's written feedback, it was found that the disagreement was related to one specific sub-category - *use of signposting.* More detailed descriptors were added to help raters to distinguish the performance at adjacent bands. For example, *good* signposting at Band 2 at ISE II indicates 'a performance with appropriate use of cohesive devices and topic sentences' and *acceptable* signposting at Band 1 indicates 'a performance with some inconsistent/faulty use of cohesive devices and topic sentences'.

For Pilot 2, the inter-rater reliability was generally high in all the rating categories across the different levels, except for *Task Fulfilment* at ISE II. A follow-up analysis was conducted to examine a subset of scripts. Some inconsistency regarding the overall communicative

effectiveness of the scripts was noted. The corresponding descriptors were refined. The inter-rater reliability at ISE III was comparatively lower than the other levels. It should be noted that the number of scripts at ISE II and ISE III in Pilot 2 was lower than the other two levels. Nevertheless, inter-rater reliability at these two levels should be monitored in the pretesting phase.

Regarding the rating scale use, the Rasch analysis showed that all the 6 raters were consistent in the use of rating scales across all the ISE tests (i.e. fit the model well, judging from the infit mean square values of 0.75 to 1.30 (Bond and Fox, 2007)). There were 3 test takers who did not quite fit the model at ISE II. It was advised that these scripts do not go into the rater training materials as the benchmark scripts, but should be used as examples of 'anomalous' cases which might need careful consideration. Also, as the raters get used to the revised rating system, it is expected that fewer 'outliers' like these scripts are likely to be observed.

Other results from Rasch analysis confirmed that the revised rating scales were working well. Tables 14 and 15 summarise the logit values of each rating category and of each band (to show the relative difficulty of a rating category compared to the others).

According to Table 11, the category of *Reading for Writing* was the most difficult (for achieving a high score) in Pilot 2. This was expected, as this task type was new and the test takers in Pilot 2 did not score very highly in this category. After the launch, publication of exam guides and more practice materials may help learners better prepare for the expectation of this criterion.

Another set of information which confirmed the appropriate use of the rating scales were those of 'rating category statistics' calculated by FACETS. One of the three key features of the category statistics to be looked at is *the average measure*, which indicates the average test taker-

ability measure represented by each level. This should increase as the level goes up (Bond & Fox, 2007). The second feature, called *threshold*, shows the lowest test taker-ability measure that a level is most likely to be assigned (Linacre, 2009). Like the average measure, this feature should increase monotonically across the levels. The third feature is the *outfit mean square* of each level which estimates the fit to the model; if it is larger than 2.0, collapsing the level to an adjacent level should be considered (Linacre, 2004). The rating category statistics are summarised in Table 15.

The average measures show a steady increase across the levels, as do the threshold values. The thresholds were more than 1.4 logits apart, which indicates that there is sufficient distance between the bands (Bond and Fox, 2007). The outfit mean-square values were less than 2.0, all of which indicate that the rating scales are functioning well and that the raters applied them in a consistent manner.

5. Conclusions

The need to develop specific rubrics to address the unique features of the integrated reading-into-writing task type seems well-established in the literature, but actual examples of such reading-into-writing rubrics are relatively rare. This paper reported an empirical mixed methods project to redevelop and validate a suite of level-specific and analytic reading-into-writing rubrics. The project developed the initial rubrics to address the prioritised needs of the ISE exams and then refined these rubrics iteratively through a series of pilots and analyses. Several limitations in the methodology should be noted. Due to practical constraints, the number of scripts in Pilot 2 at the four levels was not as comparable as might be desirable. Only a small set

of automated textual indices was used to analyse the complexity of the scripts, and these indices might not be most effective in differentiating the distinctive features at different levels. The findings were helpful in supplementing the qualitative expert panel's analysis, but a more substantial automated textual analysis may provide valuable information at the pre-testing or live launch stages. The development and validation project was able to provide validity evidence that the rubrics met key quality standards recognised by the language testing industry and were ready for pre-testing (i.e. the last stage before live launch). The approach outlined in this project will hopefully benefit not only other language testers with an interest in developing reading-into-writing rubrics, but also other colleagues in more localised testing settings.

On a final note, it should be emphasised that the rubrics are a part of the whole cycle of test development and validation. No rubrics stand alone; the analytic rubrics reported in this paper are designed to be applicable to performances on the tasks that are designed based on test specifications, be accompanied by rater training (and monitoring) with good benchmark and practice materials, and be informative in giving useful feedback to test takers and their teachers. The descriptors are sufficiently generic to allow variations within different testlets, but are still specific to the characteristics of the performance elicited by this task type specified in the ISE test specifications. The 'generic' descriptors, therefore, need to be complemented by initial rater training and ongoing re-certifying which should familiarise raters with actual source texts and tasks, so as to enable them to tailor their judgements to the requirements. While this exercise is both time- and resource-consuming, moving the rating approach from holistic to analytic rubrics is believed to bring further positive washback in student's learning. For instance, teachers will have more support to provide diagnostic feedback on students' integrated writing performance, as well as to identify the areas that their students need further training in. As stated earlier, since

ISE is intended for young people or adults in education, the move towards analytic rating scales has been essential for enhancing its significant feature for providing detailed feedback which will facilitate learning.

ISE is currently in the first year of its launch. It is hoped that, as more test data, test taker performance, scores, rating processes and feedback are compiled, further evidence of validity will be sought as part of ongoing validation of the exams.

References

AUTHOR, DATE

AUTHORS, DATE

Bond, T.G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the Human Sciences*. (2nd ed.) Mahwah, N.J.: Erlbaum.

Cobb, T. (2003). VocabProfile, The Compleat Lexical Tutor. Retrieved from http://www.lextotor.ca.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching and assessment.* Cambridge: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing*, *21*(2), 107–145.

Cumming, A. H., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a*

preliminary analytic framework. (TOEFL Monograph No. MS-22). Princeton, NJ: Educational Testing Service.

ETS (2004). *TOEFL iBT scoring guides (Rubrics) for writing responses*. Retrieved from https://www.ets.org/Media/Tests/TOEFL/pdf/Writing_Rubrics.pdf.

Fulcher, G. (1996). Does thick description lead to smart tests? A rating-scale approach to language test construction. *Language Testing, 13*(2), 208-238.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Longman/Pearson Education Ltd.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507-531.

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spaan Fellow working papers in second/foreign language assessment, 7,* 47-84. Ann Arbor: The University of Michigan.

Gebril, A. & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing, 21*, 56-73.

Green, A. (1998). *Verbal protocol analysis in language testing research: a handbook.* Studies in Language Testing, 5. Cambridge: Cambridge University Press.

Green, A. B. (2012). *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range*. Cambridge: Cambridge University Press.

Green, A. B. (2013). *Trinity ISE testing constructs project: A review and recommendations for development.* Report submitted to Trinity College London.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Rowley, MA: Newbury House.

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics, 25,* 205–227.

Lim, G. (2012) Developing and validating a mark scheme for writing. *Cambridge English Research Notes, 49,* 6-10.

Linacre J. M. (2004). Predicting Measures from Rating Scale or Partial Credit Categories for Samples and Individuals. *Rasch Measurement Transactions, 18*(1) 972.

Linacre J. M. (2009). Unidimensional models in a multidimensional world. *Rasch Measurement Transactions. 23*(2), 1209.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

North, B. (2000). *The development of a common framework scale of descriptors of language proficiency based on a theory of measurement*. PhD thesis, Thames Valley University/New York: Peter Lang.

Plakans, L. M. & Gebril, A. (2012). A close investigation into source use in L2 integrated writing tasks. *Assessing Writing, 17*(1), 18-34.

Shaw, S. D. & Weir, C. J. (2007). *Examining writing: research and practice in assessing second language writing.* Cambridge: UCLES/Cambridge University Press.

Shohamy, E. (1990). Discourse analysis in language testing. *Annual Review of Applied Linguistics*, *11*, 115-31.

Trinity College London. (2015a). *Integrated Skills in English (ISE) Specifications. Reading & Writing. ISE Foundation to ISE III*. Retrieved from http://www.trinitycollege.com/site/?id=3192.

Trinity College London. (2015b). *Integrated Skills in English (ISE) – revised*. Retrieved from http://www.trinitycollege.com/site/?id=3193.

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

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*(1), 27-55.

Wu, R. Y. F. (2014). *Validating second language reading examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference*. Studies in Language Testing 41. Cambridge: Cambridge University Press.

Acknowledgement

The project reported in this paper was commissioned to the research team by Trinity London College. The authors express their gratitude to the two anonymous reviewers for their constructive reviews. They would also like to thank Nicola Latima for her valuable assistance in the project.

Table 1

Prioritised recommendations for the ISE rubrics development project

- Develop new descriptors to differentiate the distinct features of performance elicited from the integrated reading-into-writing task and the independent writing task
- Change from a holistic to an analytic approach to scoring to enable score reporting for diagnostic purposes and bring positive washback effect in teaching and learning
- Change from generic performance descriptors to level-specific performance descriptors for the individual levels of the ISE suite to better reflect test takers' performance requirements at each level
- Reduce the number of bands within each ISE level to four bands
- Remove the checklist of language forms from the rubrics to avoid negative washback of teaching and learning in the classroom

Table 2

Timeline for developing the rubrics: May 2013 to January 2014

- Initial project meeting with Trinity's ISE redevelopment team to set goals for the redevelopment project
- Development of test specifications for the ISE R/W Exam suite at four levels: ISE F, ISE I, ISE II and ISE III
- Development of prototype reading-into-writing tasks at each ISE level
- Mini-pilot (n=40) and analysis of test takers' performance
- Development of draft rubrics at each ISE level
- Revision of test specifications and prototype test papers
- Development of rater training materials by Trinity
- Rater training pilot 1 (n=4)
- Pilot 1 (n=186), analysis of test takers' performance and raters' group interview (n=4)
- Revision of test specifications, prototype test papers and rubrics
- Rater training pilot 2 (n=12)
- Pilot 2 (n=416), raters' feedback questionnaire (n=12)
- Finalisation of test specifications, prototype test papers and rubrics

Table 3

Outline of the reading-into-writing tasks

Level	Task outline	Target processes
ISE F	Write an essay of 70-100 words using factual information from three short texts (including two straightforward factual descriptive texts and one non-verbal input, e.g. diagram, figure) with a total of 300 words	• identify factual information that is relevant to the writing task across multiple texts; • paraphrase/summarise key words and phrases or short sentences; and • incorporate such information to produce a short and simple response to suit the purpose for writing (e.g. to provide a solution to a straightforward problem)
ISE I	Write an essay of 100-130 words using information from four short texts (including three straightforward factual descriptive texts and one non-verbal input, e.g. diagram, figure) with a total of 400 words	• identify straightforward information that is relevant to the writing task and the main conclusions, significant points and common themes across multiple texts; • paraphrase/summarise short pieces of information; and • incorporate such information to produce a short and simple response to suit the purpose for writing
ISE II	Write an essay of 120-170 words based on four short texts (including three texts with factual ideas, opinions, argument or discussion and one 1 non-verbal input, e.g. diagram, figure) with a total of 500 words	• identify information that is relevant to the writing task and the common themes and links across multiple texts; • paraphrase/summarise factual ideas, opinions, argument and/or discussion; and • synthesise such information to produce coherent responses to suit the purpose for writing (e.g. to offer solutions to a problem and/or evaluation of the ideas)
ISE III	Write an essay of 170-220 words based on four texts (including three texts with information, ideas or opinions at detail level and one non-verbal input, e.g. diagram, figure) with a total of 700 words	• identify information that is relevant to the writing task and the common themes and links across multiple texts and the finer points of details, e.g. attitudes implied; • paraphrase/summarise complex and demanding texts; and • synthesise such information to produce elaborated responses with clarity and precision

Table 4

Number of test-takers

	ISEF (A2)	ISEI (B1)	ISEII (B2)	ISEIII (C1)
Mini pilot	10	10	10	10
Pilot 1	45	58	34	49
Pilot 2	129	183	52	52

Table 5

Automated textual indices

Automated textual indices	**Definition**
Total text length	The total number of words per text
High frequency words (K1)	The ratio of words which appear in the first most frequent 1000 BNC (2001) wordlist to the total number of words per text
High frequency words (K2)	The ratio of words which appear in the second most frequent 1000 BNC (2001) wordlist to the total number of words per text
Academic words	The ratio of words which appear in the Academic Wordlist (Coxhead, 1998) to the total number of words per text
Low frequency words (Off-list)	The ratio of words that do not appear in either the most frequent 15,000 BNC wordlist to the total number of words per text
Type-token ratio	The number of unique words divided by the number of tokens of these words
Lexical density	The number of content words divided by the total number of words

Table 6

The analytic assessment criteria and sub-categories (excluding descriptors)

READING FOR WRITING	**TASK FULFILMENT**	**ORGANISATION AND STRUCTURE**	**LANGUAGE CONTROL**
i. understanding of source materials	*i. overall achievement of communicative aim*	*i. text organization, including use of paragraphing, beginnings / endings*	*i. range and accuracy of grammar*
ii. selection of relevant content from source texts	*ii. awareness of the writer-reader relationship (style and register)*	*ii. presentation of ideas and arguments, including clarity and coherence of their development*	*ii. range and accuracy of lexis*
iii. ability to identify common themes and links within and across the multiple texts	*iii. adequacy of task coverage*	*iii. consistent use of format to suit the task*	*iii. effect of linguistic errors on understanding*
iv. adaptation of content to suit the purpose for writing		*iv. use of signposting*	*iv. control of punctuation and spelling*
v. use of paraphrasing and/or summarising			

Table 7

Examples of expert panel references to features of the *Reading for Writing* criterion at each ISE level

Reading for Writing	ISE F	ISE I	ISE II	ISE III
Understanding of source materials	*misinterpreted some of the content** (001)	*showed very limited understanding of source materials** (***102***)	*some misinterpretation of the sources** (243)*; showed good but not full understanding of the source text*** (205)	*showed a good understanding of the passages including more challenging implicit meaning*** (355)
Selection of relevant content from source texts	*all information selected is relevant but some necessary points are missing** (010)		*seemed to make use of one source text only** (243); *included some irrelevant ideas** (238)	*most of the source text ideas were irrelevant** (362)
Ability to identify common themes and links within and across the multiple texts	*able to show how different ideas relate to each other ** (008)	*did not address the common theme of the source texts** (124); *able to make the connections between the source texts*** (***118***)	*grabbed the common theme of the source texts well and was able to discern overlapping ideas*** (236)	*showed good intertextual reading ability to identify links between different texts*** (341)
Adaptation of content to suit the purpose for writing	*pulled relevant ideas together but did not really use them for the purpose of the writing** (001); *adapted the information from source texts well for the purpose of writing*** (011)	*did not really adapt the content for the purpose of writing** (***102***); *adapted the content well for the writing task*** (***118***)	*did not use the source texts to offer solutions* (208); very good transformation of source texts to suit the purpose of writing*** (236)	*attempted to adapt the source materials for own writing but struggled to reorganise them for the writing*** (338) *good evaluation of the source text ideas based on the purpose for writing*** (341)
Use of paraphrasing and/or summarising	*very poor paraphrasing skills** (003); *good summarising skills*** (011)	*heavy lifting** (101); *good summarising skill*s** (***122***)	*demonstrate hardly any evidence of paraphrasing or summarising skills** (208); *acceptable paraphrasing and summarising skills at this level*** (205)	*evident of heavy lifting** (344); *fairly pretty good summarising skills of the complex texts*** (355)

( ) indicates the reference number of the scripts and the ***bold italics*** indicates the scripts found in Appendix 3.

*weaker scripts

**stronger scripts

Table 8
Comparisons of task response length at each ISE level

Levels	Mean	Std. Dev
ISE F	93.37	20.50
ISE I	147.60	43.84
ISE II	196.95	67.68
ISE III	239.70	59.08

Table 9
Automated textual analysis of the written scripts

		Mean	**Std. Dev**
K1	ISEF	84.12	5.25
	ISEI	80.26	4.83
	ISEII	83.51	4.07
	ISEIII	77.91	3.65
K2	ISEF	84.62	20.12
	ISEI	83.39	19.19
	ISEII	90.18	3.07
	ISEIII	82.79	3.26
AWL	ISEF	1.48	1.49
	ISEI	3.05	2.22
	ISEII	5.02	2.59
	ISEIII	3.47	1.57
Off-list	ISEF	9.30	4.05
	ISEI	9.34	3.37
	ISEII	4.80	2.98
	ISEIII	13.74	2.91
Type-token ratio	ISEF	0.65	0.08
	ISEI	0.59	0.06
	ISEII	0.57	0.05
	ISEIII	0.55	0.04
Lexical density	ISEF	0.55	0.05
	ISEI	0.53	0.08
	ISEII	0.49	0.11
	ISEIII	0.51	0.04

Table 10

Summary of raters' discussion of the differences between the *Reading for Writing* and *Task Fulfilment* criteria

Reading for Writing	**Focus**	**Task Fulfilment**	**Focus**
Understanding of source materials	This is to determine whether the candidate understood the source materials correctly. Some candidates showed obvious misunderstanding.	***Overall achievement of communicative aim***	This is to determine whether the candidates were able to achieve the overall communicative purpose of the task
Selection of relevant content from source texts	This is to determine whether the candidates were able to select relevant content from different source texts, and whether they were able to identify irrelevant content. Some candidates included ideas from one source only and some students included irrelevant contents.	***Awareness of the writer-reader relationship (style and register)***	This is to determine whether the candidates had a good awareness of the need of the readers by writing in an appropriate style and register.
Ability to identify common themes and links within and across the multiple texts	This is to determine whether candidates were able to work out the intertextual macro-structures. At lower levels, candidates could discern when the same idea has been mentioned in more than one text. At higher levels, candidates showed ability to create intertextual hierarchy.	***Adequacy of task coverage***	This is to determine whether the writing has met the task requirements (e.g. genre, topic and number of words)
Adaptation of content to suit the purpose for writing	This is to determine whether candidates were able to adapt the content of the source texts appropriately, e.g. to offer solutions, to evaluate the ideas, etc.		
Use of paraphrasing and/or summarising	This is to determine whether candidates were able to paraphrase, summarise and/or direct quote source materials as appropriate to their level.		

Table 11
Raters' feedback questionnaire in Pilot 2

		% of raters who strongly agreed/ agreed to the statement
1	The descriptors were easy to understand and interpret in the category of:	
	• Reading for writing	75.0
	• Task Fulfilment	83.3
	• Organisation and Structure	91.7
	• Language Control	83.3
2	The descriptors distinguished well between each of the bands (i.e. Bands 0, 1, 2, 3) in the category of:	
	• Reading for writing	75.0
	• Task Fulfilment	91.7
	• Organisation and Structure	91.7
	• Language Control	83.3
3	Each category was distinct from the other three categories.	75.0
4	The written scripts generally provide a sufficient quantity/quality of language to rate appropriately.	91.7
5	The descriptors for each band are appropriate.	100

Table 12
Inter-rater reliability (Pilot 1)

	Reading for Writing	**Task Fulfilment**	**Organisation & Structure**	**Language Control**
ISE F	0.80	0.83	0.83	0.83
ISE I	0.85	0.83	0.89	0.89
ISE II	0.86	0.82	0.59	0.86
ISE III	0.84	0.84	0.86	0.83

Table 13
Inter-rater reliability (Pilot 2)

	Reading-to-Write	**Task Fulfilment**	**Organisation & Structure**	**Language Control**
ISE F	0.89	0.88	0.83	0.83
ISE I	0.86	0.78	0.78	0.80
ISE II	0.84	0.59	0.89	0.91
ISE III	0.71	0.80	0.75	0.73

Table 14

Logit values of each rating category (Pilot 2)

Analytic Categories	**Measure (logit)**	**S.E.**
Organisation	-0.84	0.19
Task Fulfilment	-0.23	0.19
Language Control	0.26	0.19
Reading for Writing	0.81	0.21

Table 15

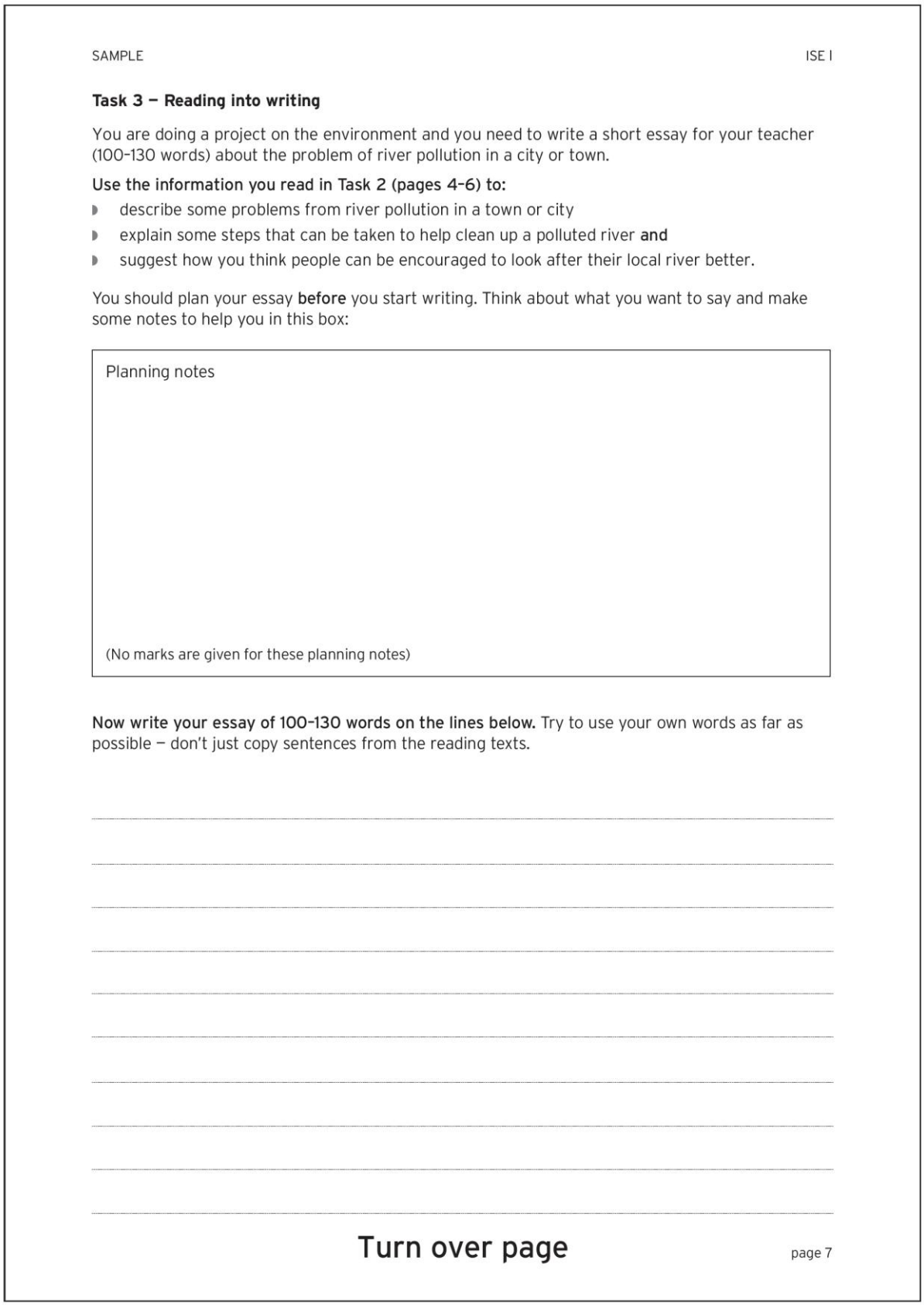
Summary of Category Statistics (Pilot 2)

Band	Average measures	Threshold	Outfit mean square
0	-4.34	None	0.9
1	-2.07	-3.63	1.1
2	0.31	-0.43	1.0
3	2.60	4.08	1.1

Figure

Appendix 1: Sample ISE I reading-into-writing task and texts (i.e. Texts A to D)

Note. Candidates go through Texts A to D in Task 2 (Reading comprehension) of the test paper.



SAMPLE ISE I

Task 3 – Reading into writing

You are doing a project on the environment and you need to write a short essay for your teacher (100–130 words) about the problem of river pollution in a city or town.

Use the information you read in Task 2 (pages 4–6) to:

- describe some problems from river pollution in a town or city
- explain some steps that can be taken to help clean up a polluted river **and**
- suggest how you think people can be encouraged to look after their local river better.

You should plan your essay **before** you start writing. Think about what you want to say and make some notes to help you in this box:

Planning notes

(No marks are given for these planning notes)

Now write your essay of 100–130 words on the lines below. Try to use your own words as far as possible – don't just copy sentences from the reading texts.

Turn over page

page 7

Text A

From: Eva
Sent: 11 November 2014 19:37
To: editor@eveningnews.co.uk
Subject: River Mêle

Dear Editor

The River Mêle causes health problems in the city, so we need to take action. Although the other local factories have stopped putting waste into the river, the paper factory is still breaking pollution laws, and should have to pay big fines.

The mud of the riverbed needs to be taken away because it's polluted with chemicals. Politicians are scared to say this, because it brings jobs to the city, but it is obvious that the paper company should pay.

Also, people need to be educated: drinks bottles and plastic bags wouldn't be such a problem if people reused or recycled them.

Yours
Eva Strauss

Text B

The River Tollen: Yearly report on the results of pollution

- Waste from factories up 6.4%
- Oxygen (O_2) levels in water down 3.5%
- Fish numbers down 3%
- Water birds down 2.4%
- Rubbish up 14%
- Chemicals from farms in riverbed up 5.5%

Text C

The city is getting millions from the government to improve the River Lamm! How should it spend the money?

Paul: I've always thought that the river would be great for kayaking so how about a watersports centre for people to do things like that?

Marcus: It would be great to see people using the river for transport, like in the old days.

Divna: Fear stops a lot of people going to the river! Even a little lighting along the riverbank would help people to feel safe.

Alex: I'd like to see one of the old factories become a museum of the city's industrial history.

Inge: @Alex And some quality waterside cafés would attract visitors too.

Simone: @Divna I agree – security cameras too, to protect people from criminals!

Alex: @Inge Hopefully they'll close that fast food place – that would mean less litter on the ground!

Text D

GREEN CITY — NEWS

In the yearly Big Clean-up on the River Vico, 50 students picked up rubbish from the banks of the river, and several local companies got together to clear the river of fridges, bikes and other large items! We criticise supermarkets on this site sometimes, but they let staff have time off work to plant trees along the river, so well done to them!

Science student Martina Keller took part in the Clean-up. She told us, 'In the five years since this started, you can see the change — the river's clear again now, not black, like it used to be! Plants are growing on the bottom of the river again, and we'll see a lot more fish and birds, I'm sure.'

Source: http://www.trinitycollege.com/site/?id=3194

Appendix 2: ISE II Reading-into-Writing Rubric

Note. The 4 bands were labelled as Bands 0, 1, 2 and 3 during the developmental and validation phase of the ISE reading-into-writing rubrics, and thus are referred to as such throughout this paper. During the pre-testing phase (which is beyond the scope of this paper), the 4 bands were renamed as Score 1 (the original band 0), 2 (the original band 1), 3 (the original band 2) and 4 (the original band 3). Score 0 was then used to distinguish scripts which do not need further rating. Here is the published version of the ISE II reading-into-writing rubric, which is currently in use in ISE exams.

Score	Reading for writing • Understanding of source materials • Selection of relevant content from source texts • Ability to identify common themes and links within and across the multiple texts • Adaptation of content to suit the purpose for writing • Use of paraphrasing/summarising	Task fulfilment • Overall achievement of communicative aim • Awareness of the writer-reader relationship (style and register) • Adequacy of topic coverage
4	• Full and accurate understanding of the essential meaning of all source materials demonstrated • A wholly appropriate and accurate selection of relevant content from the source texts • Excellent ability to identify common themes and links within and across the multiple texts and the writers' stances • An excellent adaptation of content to suit the purpose for writing • Excellent paraphrasing/summarising skills of factual ideas, opinions, argument and/or discussion demonstrated	• Excellent achievement of the communicative aim • Excellent awareness of the writer-reader relationship (ie appropriate use of standard style and register throughout the text) • All requirements (ie genre, topic, reader, purpose and number of words) of the instruction appropriately met
3	• Full and accurate understanding of the essential meaning of most source materials demonstrated • An appropriate and accurate selection of relevant content from the source texts (ie most relevant ideas are selected and most ideas selected are relevant) • Good ability to identify common themes and links within and across the multiple texts and the writers' stances • A good adaptation of content to suit the purpose for writing (eg apply the content of the source texts appropriately to offer solutions, offer some evaluation of the ideas based on the purpose for writing) • Good paraphrasing/summarising skills of factual ideas, opinions, argument and/or discussion demonstrated (with very limited lifting and a few disconnected ideas)	• Good achievement of the communicative aim (ie easy to follow and convincing for reader) • Good awareness of the writer-reader relationship (ie appropriate use of standard style and register throughout the text) • Most requirements (ie, genre, topic, reader, purpose and number of words) of the instruction appropriately met
2	• Full and accurate understanding of more than half of the source materials demonstrated • An acceptable selection of relevant content from the source texts (the content selected must come from more than one text) • Acceptable ability to identify common themes and links within and across the multiple texts and the writers' stances (eg ability to discern when the same idea has been mentioned in several texts and therefore avoid repeating it) • Acceptable adaptation of content to suit the purpose for writing • Acceptable paraphrasing/summarising skills of factual ideas, opinions, argument and/or discussion demonstrated	• Acceptable achievement of the communicative aim • Some awareness of the writer-reader relationship • Most requirements (ie genre, topic, reader, purpose and number of words) of the instruction acceptably met
1	• Inaccurate and limited understanding of most source materials • Inadequate and inaccurate selection of relevant content from the source texts (ie fewer than half of the relevant ideas are selected and most of the selected ideas are irrelevant) • Poor ability to identify common themes and links within and across the multiple texts and the writers' stances (ie misunderstanding of the common themes and links is evident) • Poor adaptation of content to suit the purpose for writing (ie does not use the source texts' content to address the purpose for writing) • Poor paraphrasing/summarising skills of factual ideas, opinions, argument and/or discussion (with heavy lifting and many disconnected ideas)	• Poor achievement of the communicative aim (ie difficult to follow and unconvincing for reader) • Poor awareness of the writer-reader relationship • Most requirements (ie genre, topic, reader, purpose and number of words) of the instruction are NOT met
0	• Task not attempted • Paper void • No performance to evaluate	

Score	Organisation and structure • Text organisation, including use of paragraphing, beginnings/endings • Presentation of ideas and arguments, including clarity and coherence of their development • Consistent use of format to suit the task • Use of signposting	Language control • Range and accuracy of grammar • Range and accuracy of lexis • Effect of linguistic errors on understanding • Control of punctuation and spelling
4	• Effective organisation of text • Very clear presentation and logical development of most ideas and arguments, with appropriate highlighting of significant points and relevant supporting detail • Appropriate format throughout the text • Effective signposting	• Wide range of grammatical items relating to the task with good level of accuracy • Wide range of lexical items relating to the task with good level of accuracy • Any errors do not impede understanding • Excellent spelling and punctuation
3	• Good organisation of text (eg appropriately organised into clear and connected paragraphs, appropriate opening and closing) • Clear presentation and logical development of most ideas and arguments, with appropriate highlighting of significant points and relevant supporting detail • Appropriate format in most of the text • Good signposting (eg appropriate use of cohesive devices and topic sentences)	• Appropriate range of grammatical items relating to the task with good level of accuracy (with mostly non-systematic errors) • Appropriate range of lexical items relating to the task with good level of accuracy (without frequent repetition • Errors only occasionally impede understanding • Good spelling and punctuation (may show some signs of first language influence)
2	• Acceptable organisation of text • Presentation and development of most ideas and arguments are acceptably clear and logical, with some highlighting of significant points and relevant supporting detail • Appropriate format in general • Acceptable signposting (eg some inconsistent/faulty use of cohesive devices and topic sentences)	• Acceptable level of grammatical accuracy and appropriacy relating to the task, though range may be restricted • Acceptable level of lexical accuracy and appropriacy relating to the task, though range may be restricted • Errors sometimes impede understanding • Acceptable spelling and punctuation
1	• Very limited or poor text organisation • Most ideas and arguments lack coherence and do not progress logically • Inappropriate format throughout the text • Poor signposting (eg inappropriate or poor use of cohesive devices and topic sentences)	• Inadequate evidence of grammatical range and accuracy (may have control over the language **below** the level) • Inadequate evidence of lexical range and accuracy (may have control over the language **below** the level) • Errors frequently impede understanding • Poor spelling and punctuation throughout
0	• Task not attempted • Paper void • No performance to evaluate	

Source: http://www.trinitycollege.com/resource/?id=6292

Appendix 3: Samples of scripts at ISE I (glossed with band ratings in the 4 rating categories)

Note. RW = Reading for Writing, TF = Task Fulfilment, OS = Organisation & Structure, LC = Language Control

Script ref. no.
102

RW	TF	OS	LC
Band 0	Band 0	Band 0	Band 0

Do you care about Pollution? So you knaw if throw things to the floor it contaminating the earth, if you care of contaminate we can help. In Uruguay they throw things to the floor and rivers like: .... fruits, vegetables plastic bag. And sometimes ho has a dog when the dog do his necesitis do you have to have a plastic bag for take it from the floor

In rivers you dind have to throw things keep it in plastic bag and then throw it in a containor. If you see throw in the river you can take it and then throw it in a garbech. If do you do this you are going to help the earth

Script ref. no.
118

RW	TF	OS	LC
Band 3	Band 2	Band 2	Band 2

Rivers - Our Second Life

Rivers are of economic importance for a country. For example River Mêle. Its overpolluted!! Its against the law to put waste into the river but its being broke by the paper factor. If the laws are made stricter and fines heavier then this can be stopped.

But not all the rivers are polluted, the Riven Vico is getting much more cleaner. Its all due to the efforts of the people. ... handwork for the past 5 years will surely make River clean and beautiful again.

We can make the place more fascinating by adding lightings to the place, it makes it very beautiful as well as it becomes a safe place. Boating, watersports all these activities improve the River.

So a little more handwork would surely make all the rivers pollution free. Make it happen and get going.!!!

Script ref. no.
122

RW	TF	OS	LC
Band 3	Band 2	Band 2	Band 2

River pollution became a big problem in this present generation. It causes many problems to the people living along the river banks such as health and sanitary problems. Many factories and industries which are built along the rivers throw all the harmful and waste chemicals into the water.

This results in the death of many aquatic plants and animals. People often throw household waste and rubbish into the water which is also causing pollution.

Every problem has solution. We can also prevent water pollution by educating the people on the problems of polluting the water. The mud mean the river buds must be taken as it consists of a lot of harmful chemicals and if along with the water, if the mud with the chemicals go into the water, then the water gets polluted. Campaigns must be conducted to educate people.

Polluting water is very bad as it is a very important source of life. Hence, we all should become aware and save our planet Earth.