# Health Sciences Communication Skills Test:

# the development of a rating scale

**Mizue Inoue**

**University of Melbourne**

## 1.        Overview of the study

*This paper attempts to refine an existing rating scale for the Health Sciences Communication Skills Test (HCST) administered at the University of Melbourne, in which some of the rating criteria are not fully defined at each proficiency level. The study examines 12 videotaped oral performances derived from operational administrations of the HCST. The samples selected to represent high, middle and low proficiency levels on the basis of scores assigned by HCST were transcribed and analysed using a range of discourse measures considered to be relevant to the test construct. The utility of each of these measures in discriminating between learners of different proficiency levels was established by comparing their relative frequency in speech samples rated as being of high, medium and low levels of proficiency. The study uses these measures to propose refinements to the current descriptors for one of the rating categories. It is argued that a discourse-based rating scale offers a more valid means of diagnosing learners' strengths and weakness as well as a more valid means of communicating test results than the traditional impressionistic scale.*

## 2. Literature Review

### 2-1 Rating scales

In language testing, a rating scale is a series of ascending descriptions of salient features of performance at each language level (McNamara, 2000). A language performance can be assessed by examining either the whole impression of the performance or the performance according to different criteria. In this regard, there are two types of rating scales: holistic scales which describe learners' performances as a whole (e.g. the American Council for the Teaching of Foreign Languages scale); and analytic scales which consist of a number of criteria referring to particular aspects of performance such as grammar, fluency and content (e.g. the Test of Spoken English).

The purpose of the rating scale determines the nature of the description. Alderson (1991) presented three types of rating scale: user-oriented scales, constructor-oriented scales and assessor-oriented scales. User-oriented scales offer useful information to test-users so that they would infer what the test-taker can do in a number of possible situations. Constructor-oriented scales provide guidance for task developers to select specific tasks in order to elicit test-takers' language samples. Assessor-oriented scales are designed to facilitate rating procedures along with bringing consistency to the process. Pollitt and Murray (1996) added diagnosis-oriented scales, which are based on assessor-oriented scales but consist of more detailed descriptions to transfer diagnostic information.

In relation to rating scale development, there are two basic approaches: intuitive methods and empirical methods (Fulcher, 2003). In the former case, a rating scale is designed by an experienced teacher(s) or test developer(s) in response to existing rating scales or the teaching syllabus (e.g. the American Council for the Teaching of Foreign Languages scale). In the latter case, actual learners' performances are analysed to identify key features at each proficiency level. A number of studies have developed and refined existing rating scales using this approach, and this will be discussed in the following

section.

## 2-2      Rating scale development using test-taker performance

In most cases existing rating scales were developed through intuitive methods and there is therefore no assurance that their level descriptions are accurate (North & Schneider, 1998). One of the approaches to tackling this issue is to examine actual performances produced by learners to understand their performance features at each level.

Upshur and Turner (1995) presented an empirically developed set of rating scales, known as the empirically derived, binary-choice, boundary definition scales (EBBs). The EBBs consist of a number of questions established by a team of experts through the   analysis of sample performances, which facilitate raters' judgements on learners' performances. Although Upshur and Turner argued the effectiveness of an empirically derived rating scale and its positive effects on the validity of ratings, the study included no analysis of the test performance itself.

Fulcher (1987) discussed the assessment scale of the English Language Testing Service (ELTS) in terms of fluency through a close examination of a conversation among four speakers under non-test condition. Even though the participants were three native speakers of English and one high-proficient non-native speaker, their performances were not consistent with the ELTS assessment criteria. Fulcher argued that the ELTS assessment scale was theory-based rather than a depiction of what happens in real communicative situations and called for empirical justification for the scales.

Fulcher (1996) made a case for the usefulness of examining actual test-taker performances as a basis for a rating scale development. His study examined 21 audio-recordings of oral interviews in relation to fluency through both qualitative and quantitative approaches. The interview data were coded into eight categories, which were identified as potential explanations for disfluency. The results showed

that different levels of test-taker performances were predicted well by most of the selected categories. The study contributed to the development of the rating scale descriptions by providing more detailed language use in terms of fluency.

A study conducted by Douglas (1994) investigated the relationship between scores on the semi-direct speaking test and the actual performances by the test-taker. Six Czechoslovakian students studying at an American university took the AGSPEAK, a screening test for potential candidates for an agribusiness exchange program. The results showed that there was little relationship between the score awarded and the features of the actual performance; that is, the scores did not differentiate well enough between levels. The researcher hypothesized that this may have been due to poor rating by raters, who were possibly influenced by different aspects of the discourse including those which did not appear in the scale.

Knoch (2007) developed a rating scale used in the Diagnostic English Language Needs Assessment (DELNA) at the University of Auckland. A number of analytic measurements were performed to analyse 601 writing performances. The only measures that discriminated between the different levels of the performances were reflected in the refined rating scale. The study demonstrated that scores based on empirically developed descriptors were more reliable since detailed descriptors derived from test-taker performances facilitated the rating process.

To summarise, the discourse of test-taker performances on language tests has been an important resource in identifying indicators of learners' language proficiency as well as the points raters pay attention to in determining test-takers' proficiency levels. It has thus been demonstrated that a close analysis of actual speech samples can contribute beneficially to the development or refinement of rating scales. This kind of analysis is particularly important when devising scales for new types of task, including integrated tasks, which have thus far received relatively little attention in language testing research.

**2-3        Research on integrated tasks**

Integrated tasks refer to a combination of reading, listening, writing and/or speaking tasks, while in many cases, tests try to separate such areas of language ability to measure (Luoma, 2004). For example, the TOEFL iBT (Educational Testing Service, 2008) contains four integrated speaking tasks with written and/or spoken input stimuli along with two independent tasks. Such tasks enhance authenticity in language use and are frequently used in task-based tests to assess test-takers' ability to deal with a certain situation, in addition to general language skill itself (Lewkowicz 1997; Luoma 2004). Furthermore, in integrated tasks a test-taker performance is less likely to be influenced by his/her own knowledge or background as the input stimulus provides the necessary information (Weir, 1990).

Weir (1990) and Lewkowicz (1997) raise concerns with the issues regarding rating in integrated tasks. For instance, the success in summarisation tasks may depend on how test-takers understand listening or reading input stimuli; that is, the performance in one skill area (such as listening or reading) affects that of another (such as speaking and writing). A number of studies have been conducted thus far to answer such issues but most of them are concerned with integrated writing tasks.

Johns & Mayes (1990) compared 80 writing summaries between college-level EFL students of high proficiency and those of low proficiency. The performances were broken down into segments called 'idea units' and it was examined whether each unit was reproduced accurately or distorted, along with whether it was copied or paraphrased. The study found that more direct copy was observed in the performance of low proficiency students while higher group students were more likely to combine ideas.

Cohen (1993) pointed out an issue regarding the rating of integrated tasks through his research on summarisation tasks concerning the effect of instructions and rater consistency. Written

summaries of 63 EFL students in Tel Aviv were assessed by four raters based on empirically derived rating keys, which were developed from summaries of nine Hebrew speakers and nine native English speakers. Cohen argued that the experts often disagreed on the necessary key points for a successful summary as the level of agreement on selecting rating keys were 80 to 85% among the raters. In addition, rating was not as consistent as expected despite the provision of the rating keys.

Yu (2007) also pointed out issues regarding rating by exploring effects on scores using two different types of templates, which were empirically derived from five English native speakers and 100 Chinese university students. The experts and the students were asked to write summaries of two English texts. Ten frequent sentences were selected from each summary as an expert template and a popular (student) template, which were then used to mark the 100 summaries of the Chinese students. The study found that the experts and the students produced qualitatively different summaries and there were only a 50% overlap of the key points, which had a significant effect on the students' scores (e.g. a higher score were observed with the template derived from the students).

The number of studies investigating tasks which integrate listening and/or reading and speaking is thus far very limited. Some of the few studies concerning speech data were conducted by Brown, Iwashita and McNamara (2005) and Iwashita, Brown, McNamara and O'Hagan (2008) and focused on integrated speaking tasks in the TOEFL iBT. The study conducted by Brown et al. (2005) consisted of two phases and investigated both rater attention and test-taker performance. In the first phase, the study investigated the focus of rater attention by means of verbal reports. Five major foci were identified in the study, namely; linguistics, resources, phonology, fluency, and content. One of the significant findings from the verbal reports was that content was a major focus when judging test-taker performance on an integrated speaking task. In addition, the raters had different views regarding what consists of an accurate response in an integrated task, which may be problematic in terms of reliability

in rating.

In the second phase, also in the study of Iwashita et al. (2008), 200 speech samples derived from the TOEFL iBT were examined. A number of performance features such as grammatical accuracy, complexity, vocabulary, pronunciation and fluency were investigated, and the data were analysed quantitatively. The results indicated that each category helped distinguish between overall levels of performance with the categories of vocabulary and fluency exercising a strong influence on results. The differences between adjacent levels were not always clear except in relation to pronunciation, fluency and the number of word tokens. In terms of content, Brown et al. (2005) applied T-units as the measure of quantity, and schematic structure as the measure of quality. The integrated task elicited a more complex structure of test-taker speech and the quality of the speech showed a clear relationship to proficiency level (as reflected in scoring on the task). These findings emphasised the importance of empirically based scale development for integrated speaking tasks, especially with regard to criteria relating to the content of the performance which is critical to the success of the tasks.

In conclusion, only a small number of research studies concerned with rating scale development have been carried in relation to integrated tasks. More studies are needed to investigate this issue, especially in terms of the content of test-taker performances, in order to explain clearly to raters how to evaluate the content of input material as reflected in test-taker performance.

**2-4       Health Sciences Communication Skills Test (HCST)**

Grove and Brown (2001) report on the development of the Health Sciences Communication Skills Test (HCST), which is a diagnostic oral test developed for first-year undergraduate students, both international and local, enrolled in the Faculty of Medicine, Dentistry and Health Sciences at the University of Melbourne. The purpose of the test is to diagnose students' language and communication skills, which are important for their university studies and their future

career. In terms of university studies, the Faculty had just introduced a new curriculum, Problem-Based Learning (PBL), where students engage in group work problem-solving tasks and presentations. Clinical studies were also established in the first year of the course, where students communicate with patients directly. As a consequence, communication skills play a key role in the students' academic success. Moreover, the fact that there are increasing numbers of international students, more than one third of the total university enrolments has raised concerns among the educational staff regarding their language and interpersonal skills. Among health professionals effective interpersonal skills are considered to be vital for effective functioning in medical contexts because these contexts require interaction with clients and colleagues regarding diagnosis and treatment.

In view of the issues mentioned above, the HCST was developed and two tasks were introduced: a summary presentation and discussion. In the first part of the test, which will be the focus of the current investigation, a test-taker summarises the content of an eight-minute tutorial discussion on ethical issues of medical experimentation on human subjects and presents for about four minutes. The audio stimulus closely resembles the  kind of input which students have to deal with in their course and in their subsequent careers, containing information and spontaneous discussion, the content of which the test-taker is required to synthesise and reformulate (Grove et al., 2001). In the second section of the HCST a discussion takes place between the test-taker and an interviewer on a selected general topic (e.g. school system, professional abilities, hobbies etc.) and the candidate is required to explain, present, and justify their opinion. The skills required for both of these tasks are relevant to both the academic and clinical context.

Test performances are assessed in terms of both linguistic and task-related criteria and a score is given respectively from 1 (extremely weak) to 6 (excellent) on a score sheet along with comments on key features of individual performance (see Appendix A for a score sheet). The linguistic features are marked holistically

across the two tasks, and are divided into aspects of language (structure, vocabulary, accuracy etc.) and speech production (pronunciation, intonation, stress etc.). On the other hand, the task-related criteria reflect the skills required for each task, as viewed by medical staff. In Task 1, the summary presentation, the performance is assessed according to its organisation, content, style, fluency, coherence and comprehension. In Task 2, the discussion, the focus is on participation, ideas, interpersonal skills, coherence, and appropriateness of register. At the test development stage, medical staff were involved in the establishment of the assessment criteria and rating scales were developed through test trials.

## 3.      Research questions

This study aims to contribute to a rating scale development for the HCST. Although the scores assigned to actual performances were corroborated by health experts using benchmark test samples (Grove et al., 2001), the current study attempts to provide the basis for more detailed descriptions of performance features than those currently provided on the rating scale. The criterion examined in this paper is 'Content', which is also part of the current rating scale shown in Figure 1. The aim is to improve the current rating scale description as the current scale contains no specific guidance for performance in this criterion. This study addresses the following research question.

**Figure 1.** Content criterion

Content: Sufficiency and appropriateness of content

- Is the overall amount of information appropriate to a summary – not too much or too little?

- Are the key points and facts presented rather than minor ones – is the selection appropriate?

1) What are the distinguishing features of content at each assessed proficiency level, and how can these be used to elaborate upon the current rating scale?

# 4.      Methodology

## 4-1      Participants

The participants are 60 students, who were the first-year students enrolled in the Faculty of Medicine, Dentistry and Health Sciences at the University of Melbourne. Their language ability failed to meet the required standard of a post-admission screening procedure and took the HCST on the 1st and the 8th of March, 2007. Even for international students however, their language proficiency was assumed to be quite high, since they had met the English language prerequisite, an overall band score of 6.5 or more in the International English Language Testing System (IELTS), to enter the faculty. Each performance was rated once by only one rater, who also acted as an interviewer in the test. The four raters (M=1, F=3) involved are native speakers of English and have post-graduate education in language teaching. In addition, they are all trained as interviewers and raters for the HCST and are re-trained on an annual basis.

## 4-2      Instruments

The data were derived from the HCST carried out in March 2007 at the University of Melbourne. Each interview was video-recorded with a full view of the test-taker and the sound of their oral production. The interviewers scored the performance concurrently on six scales and wrote comments on each performance.

As for the test procedure, firstly the candidates are told to read information about the two tasks along with a list of vocabulary which appears in the audio material but may not be familiar to the test-takers. Next, they listen to the conversation once over the headphones and take notes as they listen. After listening, the test-takers have two minutes to prepare for the summary presentation before going into another room where the assessment is conducted.

**4-3          Procedures**

*4-3-1          Data Selection*

For the purpose of this study, twelve sets of data were selected from a pool of 60 based on the following procedure. The raw scores assigned by raters for content criteria of Task 1 (the summary presentation) were looked at to select sample data from the larger data set at high (mean 5-6), middle (mean 4-5), and low (mean 2-3) proficiency levels.

**Table 1. Summary of selected samples.**

| [Assigned level] | Task 1 Presentation | Rater |
|---|:---:|:---:|
| Test-takers (Gender) | Content | |
| *[Low]* | | |
| S1   (M) | 2 | A |
| S2   (F) | 3 | B |
| S3   (M) | 3 | C |
| S4   (F) | 3 | C |
| *[Middle]* | | |
| S5   (M) | 4 | A |
| S6   (F) | 4 | B |
| S7   (F) | 4 | C |
| S8   (M) | 4 | D |
| *[High]* | | |
| S9   (F) | 6 | A |
| S10(M) | 6 | B |
| S11(M) | 5 | C |
| S12(F) | 6 | D |

It was hoped that selecting data at each of these three levels would make it possible to identify those features which distinguish between

candidates of different ability. Four speech samples at each of the three levels, amounting to twelve samples in total (M=6, F=6) were selected randomly. The raw scores of the sampled data are shown in Table 1 along with the raters involved in the rating procedure, identified with the letters A, B, C and D.

*4-3-2     Data Preparation*

The first part of the interview was transcribed from the audio-taped data, which began just after the candidate's opening statement (e.g. "I'm starting now.") or with a candidate's utterance followed by the interviewer's cue (e.g. "Could you summarise what you heard?"). The end of the data was distinguished by the candidate's closing statement (e.g. "That's all."), or with a silence followed by the interviewer's question asking for the candidate's personal opinion (e.g. "Do you think it still happens?"). This type of question serves as a bridge into the second half of the test, the discussion part, and therefore cannot be considered a part of the summarisation task. On the other hand, interviewers are instructed to ask questions if the summary is short and lacking in detail. Hence, the speech stimulated by an interviewer's question to elicit more detailed information from the candidate (e.g. "What happened in that case?") were treated as a continuing part of the summarisation task and included in the data.

*4-4       Data Analysis*

In the following section, the rules for data transcription and coding are described.

*4-4-1     AS-unit*

The transcribed data were examined for content components applying the AS-unit as a unit of analysis. The AS-unit is designed specially for the analysis of oral speech data and, compared to other units (e.g. T-units), is more relevant to the segmentation of the oral data which is not as organized as written language (Foster, Tonkyn & Wigglesworth, 2000). The AS-unit consists of an independent clause or sub-clausal unit, as well as a subordinate clause. An independent

clause is a clause with a finite verb, such as the following example.

1.  |they test sixteen babies from baby's home| (1 AS-unit, 1clause)

2.  |the doctor didn't tell her that| (1 AS-unit, 1clause)

3.  |this disease usually would develop into cancer| (1 AS-unit, 1clause)

An independent sub-clausal unit is a phrase(s) or minor utterance. Although the utterance does not always contain a finite verb, it is thought to be an AS-unit as long as the missing component(s) can be reconstructed as seen in the example below. Neither of the utterances below contains a finite verb, but both can be elaborated as a full clause if the missing word ('were' for example 4 and 'was' for example 5) is inserted. Therefore, both utterances are considered to be an AS-unit.

4.  |The sixteen babies there| (1 clause, 1 AS-unit)

5.  |Next for the second case in {er} in New Zealand in ninety sixties| (1 clause, 1 AS-unit)

A subordinate clause(s) contains either a finite or non-finite verb with at least one other clause element, and is attached to the independent clauses without noticeable pause. It functions as a subject, verb complementation, phrasal post-modifier or complement as shown below.

6.  |It's better :: to give them the vaccination| (2 clause, 1 AS-unit)

7.  |I think :: he didn't tell anyone about that| (2 clause, 1 AS-unit)

8.  |this doctor had a vaccine :: that he wants :: to test on| (3 clause, 1 AS-unit)

9.      |I'll talk about a case :: that Andrew {er} had he studied| (2 clause, 1 AS-unit)

In the case of a coordination of verb phrase, it is considered to be a subordinate clause which belongs to the same AS-unit except where there is a change in intonation followed by a noticeable pause (Foster et al., 2000). Therefore, pauses are taken into account to decide whether the clause belongs to the same AS-unit.  For the current study, however, no special equipment was employed to measure pause length since, as Foster et al. (2000) explained, a pause of less than half second can be detected by simply listening to the audiotape. In the following examples, the speech data was coded as the same AS-unit where there was no pause before "and" (as shown in example 10), while it was coded as a different AS-unit where there was an observable pause before "and" (as shown in example 11).

10.     |and happened in the sixties and seventies :: where a woman :: diagnosed with carcinoma in situ of kiss :: and proceeded a treatment of hysterectomy| (4 clause, 1 AS-unit)

11.     |so {she} he made her :: to come twice a year| **and** diagnose her :: if it has develop to cancer or not| (4 clause, 2 AS-unit)

Inaudible utterances, one-word minor utterances (e.g. "ok."), and fillers (e.g. "er") and other phenomena, including disfluency phenomena listed by Foster et al. (2000) (e.g. false start, repetition, self-correction etc.) were excluded from the analysis. Table 2 contains an explanation of each phenomenon and how they were handled in this study accompanied by some examples, which were emphasised in bold.

**Table 2. Summary of the phenomena excluded from analysis.**

| Phenomenon | Explanation (example) | How it was dealt with |
|---|---|---|
| *Inaudible utterance* | utterance which cannot be transcribed due to the volume or ambiguity of the speech | not included in the transcription |
| *One-word minor utterance* | utterance which appears frequently but does not convey meaning | considered to be an indication of the end of the unit and excluded from the analysis |
| | (e.g.) ⎮so the issue here is :: {the} whether the baby is in high risk or not {**ok**}⎮ | |
| *Fillers* | utterance or sound which occurs while the speaker is searching for words or processing his/her idea | excluded from the analysis |
| | (e.g.) ⎮it's about {**er**} the vaccination :: done in the {**er**} babies room :: which includes sixteen baby⎮ | |
| *False start* | utterance which is produced and then abandoned or reformulated by the speaker | excluded from the analysis |
| | (e.g.) ⎮{**and the doctor who diagnosed it oh erm**} usually the treatment is to have hysterectomy⎮ | |

**Table 2. (continued) Summary of the phenomena excluded from analysis.**

| *Phenomenon* | *Explanation (example)* | *How it was dealt with* |
|---|---|---|
| *Repetition* | utterance in which the speaker repeats previously produced speech | the last utterance is applied unless it is emphasised intentionally |
| | (e.g.) ∣{**the**} the video {**that I have that I have that I have** } that I have had just now :: is about {**the**} the research done {**on the**} on the human subject∣ | |
| *Self-correction* | utterance which the speaker reformulates or utterance which was withdrawn by the candidate | excluded from the analysis |
| | (e.g.) ∣and {erm} first of all they discussed :: {**whether the risks and benefits of the experiment whether the risk will justify er I mean**} whether the benefit will justify the risk∣ | |
| | (e.g.) ∣{mmm} so this woman seemed to be lucky {**because er er I'm sorry I'm stuck**}∣ | |
| *Interruption and scaffolding* | interviewer's utterance which interrupts the speakers utterance. | the interviewer's utterance was excluded from the analysis, and the interrupted candidate's utterance was integrated into one AS-unit |
| | (e.g.) ∣and Jos answer that {erm it's more pf er}(**sorry I interrupted you there**) {that's ok so that more ah} it's more of a risk thing∣ | |

**Table 2. (continued) Summary of the phenomena excluded from analysis.**

| Phenomenon | Explanation (example) | How it was dealt with |
|---|---|---|
| *Self-talk* | utterance which occurs while the speaker is searching for the information from his/her notes or the speaker is confirming what the interviewer has asked. | excluded from the analysis |
| | (e.g.) ∣the tape {er} was discussion between the teacher {**so what's the name? er let me see**} Lis and two students, Andrew and Jos∣ | |
| | (e.g.) (what happened in that case?) ∣{**what happened in that case?**}because after the fifteen years that woman that undergo the treatment for the research she had carcinoma in situ at first∣ | |
| *Clarification* | utterance by which the speaker inquires about the task | excluded from the analysis |
| | (e.g.) ∣it's about vaccination case {**er do I need do explain the case (not too much) ok**}∣ | |

| | | | | |
|---|---|---|---|---|
| ∣ ∣ | the boundaries of AS-units utterance excluded from analysis | { } | the | candidate's |
| :: | the boundaries of clauses utterance excluded from analysis | ( ) | the | interviewer's |

*4-4-3 Content*

In relation to the Research Question, aimed at identifying the distinguishing features of the content of each speech sample at each assessed level, the data were examined in terms of both its quantity and quality. For the quantitative analysis, the number of AS-units and clauses in the transcribed data were examined. In order to investigate the quality of the speech sample, the schematic structure of the transcribed data was examined following the method demonstrated by Brown et al (2005). Seven schemata were employed to analyse the speech samples, inspired by the work by Eggins & Slade (1997), who demonstrated the notion of a schema to identify different types of speech and its characteristics and presented different types of genre. The schemata applied to this study are shown below with their definitions and examples, along with the number of clauses and AS-units in brackets.

a) Abstract

Units which provide a summary of the contents. This includes clarification of the input stimulus, the theme of the discussion and the main issues discussed.

12.    ｜according to the tutorial discussion on the case of {er} the moral issues involve in medical researches :: involving human subjects｜ (2 clauses, 1 AS-unit)

13.    ｜the tape is a discussion between the teacher {so what's the name er let me see} Lis and two students Andrew and Jos｜ (1 clause, 1 AS-unit)

b) Orientation

Units which provide information of the context of the example. This includes, in the case of this task, venue details, year, purpose of the experiment and subjects involved.

14.   |{er,} it's about a case in Melbourne :: happened in 1947|
      (2 clause, 1 AS-unit)

15.   |and {er} what happens was 16 young kids from baby's
      home|(1 clause, 1 AS-unit)

c) Events

Units which provide description of the sequence of the
experiment. This includes chronological information,
descriptions of actions taken and the number of subjects.

16.   |They've been given first time the dose was small| (1
      clause, 1 AS-unit)

17.   |Then after a few days or weeks {er} he repeated the
      test :: to see :: whether the vaccine works| (3 clauses, 1
      AS-unit)

d) Outcomes

Units which explain the consequence of the treatment. This
includes descriptions of reactions to the treatment received and
indications of the success or failure of the treatment .

18.   |and nothing happened| (1 clause, 1 AS-unit)

19.   |and all the 16 of the children got herpes | (1 clause, 1
      AS-units)

e) Issues

Units which raise issues involved in the experiment. This
includes discussion of problems, ethical issues discussed in the
input stimuli as well as the test-takers' own interpretations.

20.   |so question comes to :: whether the risk of having those
      healthy kids :: been given the vaccination :: is it justifiable

for the benefits of the health medical research⎪ (4 clauses, 1 AS-unit)

21.  ⎪and all of that {er} I think :: he didn't tell anyone about that (1 clause, 1 AS-unit)

f) Coda

Units which summarise the whole discussion. This includes the summary of the main issues. This move is optional as it does not appear consistently in the speech sample.

22.  ⎪and issue of this two case is {er year er the er the balance between…} the first case the balance between the risk and the benefit {of} of experiment and treatment ⎪ and the second one is about the informed consent⎪ (2 clause, 2 AS-unit)

The schematic structure of the discourse was developed by examining transcribed speech samples. AS-units were applied as a unit of analysis. In the process of coding, procedural utterances, such as "that's it" and "that's the first case", were excluded from the analysis. The results will be presented descriptively. In addition, the quantitative analysis will be presented using the number of AS-units and clauses for each schema and Chi-squares will be examined to ascertain whether there is a significant difference across assessed levels.

## 5. Results

In order to obtain inter-coder reliability on AS-units and schematic analysis, approximately 10% of the data were coded by a second coder, who acknowledged the primal theory of AS-units and was given instructions for schematic analysis. Inter-coder reliability was calculated using the formula suggested by Miles & Huberman (1994) and it was achieved in 75% of AS-units and 88% of schematic analysis.

**Table 3. Mean and standard deviation.**

| Levels | The number of AS-units | | The number of clauses | |
|---|---|---|---|---|
| | M | SD | M | SD |
| Low | 16.75 | 4.573 | 32.25 | 11.871 |
| Middle | 31.00 | 5.228 | 67.25 | 2.872 |
| High | 32.75 | 7.136 | 74.50 | 20.632 |
| Total | 26.83 | 9.114 | 58.00 | 11.790 |

The research question examines the distinguishing features of content at each assessed level for the purpose of developing detailed descriptors for the content criteria. The quantity of the content was examined using the number of AS-units and clauses. As the result shows in Table 3, the middle and high level speakers produced almost twice the amount of AS-units as the low level group did. However, these differences fell just short of statistical significance ($X^2$ (2) = 5.74, p = .056). On the other hand, Chi-squares analysis revealed significant differences in terms of the number of clauses ($X^2$ (2) = 17.60, p = .0001). The low group was significantly different from both the middle and high groups.

In relation to the quality of the content, the analysis revealed a slightly different schematic structure at each level. Overall, most of the speech data showed a similar structure to that of the input stimuli, which commenced with the theme of the discussion, followed by two examples of medical experiments presented by two different students. Figure 2 shows an expected schematic structure of the summary, which begins with Abstract, followed by two cases containing Orientation, Event, Outcome and Issues respectively, and ends with Coda as an optional move.

**Figure 2. Expected schematic structure of the summary.**

| | |
|---|---|
| Abstract | : the theme of the discussion, the main issues discussed |
| 1st case | |
| Orientation | : the venue, year, purpose of the experiment, the subject involved |
| Event | : chronological information, descriptions of each action, the number of the subject |
| Outcome | : the reaction to the treatment, the indication of the success or failure of the treatment |
| Issues | : the discussion of the ethical issues |
| 2nd case | |
| Orientation | : the venue, year, purpose of the experiment, the subject involved |
| Event | : chronological information, descriptions of each action, the number of the subject |
| Outcome | : the reaction to the treatment, the indication of the success or failure of the treatment |
| Issues | : the discussion of the ethical issues |
| (Coda) | : the summary of the main issues. This move is optional. |

The schematic structure derived from the low level speakers, as shown below in Figure 3, had less structure than the expected schematic structure. Although an outline of both cases was presented, the Issues, in which test-takers were supposed to present problematic points of each case, were likely to be left out from most of the speech data. In addition, there were some incidents where omitted information was elicited by the interviewer's questions as seen in

Low S3 and S4's structure in Figure 3. In Extract 1, which follows, a candidate from the low level group provided a brief background of the first case in the Orientation. In the Event and Outcome, the candidate only presented the very first stage of the experiment but failed to present the whole picture of the case, in consequence providing an inaccurate summary. In addition, the summary of the first case did not develop to the extent of discussing the problematic aspects of the experiment.

**Figure 3. Schematic structure (low level group).**

| S1 | S2 | S3 | S4 |
|---|---|---|---|
| Abstract | Abstract | Abstract | Abstract |
| 1st case | 1st case | 1st case | 1st case |
|     Orientation |     Orientation |     Orientation |     Orientation |
| |     Issue |     Event |     Event |
| | |     Outcome |     Outcome |
| 2nd case | 2nd case | 2nd case | 2nd case |
|     Orientation |     Orientation |     Orientation |     Orientation |
|     Event |     Event |     Issue |     Issue |
|     Outcome |     Outcome |     *Event |     *Event |
| |     Issue |     *Outcome |     *Outcome |
| | | |     *Event |
| | | |     *Outcome |
| | | |     *Issue |

**\* elicited by interviewer's questions**

**Extract 1. Low (test-taker S3, interviewer/rater C, score given 3).**

1st case

| | |
|---|---|
| Orientation | \|so {they} they come up with {er} two examples :: which is one happened in Melbourne in ninety forty-seven :: where babies were {er} affected by a disease :: {er} called herpes {sim} simplex\| |
| Orientation | \|{mmm} this disease {mmm} have been {er} epidemic disease\| |
| Orientation | \|{mmm} because {mmm the} the fact is :: they are living in closed environment\| |
| Orientation | \|so the babies are connected each other\| |
| Orientation | \|{mmm} in this {er} in this case {they are they had they can sixteen kids to be mmm er} sixteen kids to be {er} investigated\| |
| Event | \|{er} and {they found they} they gave {mmm} the vaccination to all the babies\| |
| Outcome | \|and {they} they found the vaccination few months later :: the babies turned to be in a good situation\| |

Overall, the middle level speech covered most of the information of the input stimuli. As seen in Figure 4, the schematic structure of the middle level speech is similar to the expected structure and contains all the schemata. Despite this, occasionally the schematic structure of the middle level group showed unusual structures. As seen in the schema of the second case in S6 and S7 (Figure 4), Orientation reappeared towards the end of the structure, abruptly providing background information before raising issues. Moreover, the description of each case was not thorough enough to present the sequential events of the case. For instance, in the Event and Outcome in the following extract (Extract 2), an outline of the first experiment

**Figure 4. Schematic structure (middle level group).**

| S5 | S6 | S7 | S8 |
|---|---|---|---|
| Abstract | Abstract | Abstract | Abstract |
| 1st case | 1st case | 1st case | 1st case |
|    Orientation |    Orientation |    Orientation |    Orientation |
|    Event |    Event |    Event |    Event |
|    Outcome |    Outcome |    Outcome |    Outcome |
|    Issue |    Issue |    Issue |    Issue |
| 2nd case | 2nd case | 2nd case | 2nd case |
|    Orientation |    Orientation |    Orientation |    Orientation |
|    Event |    Event |    Event |    Event |
|    Outcome |    Outcome |    Outcome |    Outcome |
|    Issue |    Event |    Issue |    Issue |
| |    Orientation |    Orientation | |
| |    Issue |    Issue | |
| Coda | Coda | | Coda |

**Extract 2. Middle (test-taker S6, interviewer/rater B, score given 4).**

1st case

Orientation   │and the first case tutoring students were talking about :: was {erm} the children {in Melbourne} in Melbourne's baby's home│

Orientation  │and the originally baby's home {was located} was ::where that herpes complex was {widely spread} wide spread already│

Orientation │so {the babies I mean} the subjects in the baby's home were healthy :: when that experiments were about to start│

Event      │but any way they gave small dose of vaccination to those sixteen babies│

**Extract 2. (continued) Middle (test-taker S6, interviewer/rater B, score given 4).**

| | |
|---|---|
| Outcome | ∣{and second second time when they came back erm I can't remember very well what they said} but eventually the experiment turned out :: to be the vaccination was failed∣ |
| Issue | ∣so in that case the students were arguing :: whether that vaccination though it failed :: was morally justifiable or not∣ |
| Issue | ∣so the boy {erm} I think :: his name was Andrew (that's right) :: Andrew was saying :: that {the babies already had a possibility I mean} they were already exposed in danger :: of contracting that disease∣ |
| Issue | ∣so people should have tried :: what they can :: even if it could fail∣ |
| Issue | ∣and {the girl I don't remember her name but the girl was saying that erm yeah} the girl was saying :: that {she shouldn't er} people shouldn't have given those uncertain vaccination to healthy kids∣ |
| Issue | ∣'cause {it's} it's considering human just as a subject for an experimentation |

was presented, but the description of every step of the experiment was absent. Nevertheless, there were certain features that differentiate this performance from that of the low group: the Orientations contained detailed descriptions of the subject including their initial health condition and their environment; and the Issue was included in the summary where the candidate discussed problematic points of the case.

The speech produced by the high level group predictably contained the most complex schematic structure of all the levels, as shown in Figure 5 below. Unlike the structures seen thus far, the Event and Outcome were subdivided into two or three sections, as seen in the schemas of S10 and S11's first cases and S9 and S11's

second cases. For example in the Extract 3, it can be seen that the speech data contained detailed information in the Event and Outcome, such as dosages given and reactions of the subject in each phase of the experiment, along with clear statements of time sequences. The summary also included the Issue, and problems of the case were argued. Another distinctive feature in the high level performances was that the candidate presented more precise background information about the case in the Orientation by stating the purpose of the experiment.

**Figure 5. Schematic structure (high level group).**

| S9 | S10 | S11 | S12 |
|---|---|---|---|
| Abstract | Abstract | Abstract | Abstract |
| 1st case | 1st case | 1st case | 1st case |
| Orientation | Orientation | Orientation | Orientation |
| Event | Event | Event | Issues |
| Outcome | Outcome | Outcome | Outcome |
| Issues | Event | Event | Issues |
|  | Outcome | Outcome |  |
|  | Event | Event |  |
|  | Outcome | Outcome |  |
|  | Issues | Issues |  |
| 2nd case | 2nd case | 2nd case | 2nd case |
| Orientation | Orientation | Orientation | Orientation |
| Event | Event | Event | Event |
| Outcome | Outcome | Outcome | Outcome |
| Event | Issues | Issues |  |
| Issues | Event | Coda | Issues |
|  | Outcome |  |  |
|  | Issues |  |  |

**Extract 3. High (test-taker S11, interviewer/rater C, score given 5).**

1<sup>st</sup> case

| | |
|---|---|
| Orientation | ∣So the first case presented {erm} was conducted in Broadmeadows in nineteen forty-seven∣ |
| Orientation | ∣and it was a research on the effectiveness of the herpes simplex vaccine∣ |
| Orientation | ∣and {sixteen chil} sixteen healthy children were actually {erm} selected from a baby's home∣ |
| Orientation | ∣and this was clamed :: to be a very good test ground as a good control sample :: because all of them healthy∣ |

**Extract 3. High (test-taker S11, interviewer/rater C, score given 5) (continued).**

| | |
|---|---|
| Event | ∣and they were given the first vaccine :: which was half dosed :: and then subsequent vaccine :: that was equivalent to full adult dose∣ |
| Outcome | ∣and the children was gone quite well∣ |
| Event | ∣{and was} and they were taken back few months later∣ |
| Event | ∣and the same {proce} procedure repeated∣ |
| Outcome | ∣so actually {er} the children responded quite badly to this second vaccine∣ |
| Event | ∣and {only nine full doses were given to} only {erm er} nine children were given the full doses the following day∣ |
| Outcome | ∣so the experiment shows :: that the vaccine didn't actually work∣ |
| Outcome | ∣and at the end of it all sixteen children contracted herpes simplex∣ |

**Extract 3. (continued) High (test-taker S11, interviewer/rater C, score given 5).**

| | |
|---|---|
| Issue | Iso from this experiment we can see :: that the risk to the children was actually greater than the benefit to them or to the whole communityI |
| Issue | Ibut one of the student argued :: that the babies were actually having very high risk :: of contracting this disease :: given their enclosed environment they are inI |
| Issue | I{er} because they are from baby's homeI |
| Issue | Ibut another student actually argued :: that {erm} these students were actually healthy from the very beginningI |
| Issue | Iso no matter what they don't actually deserved :: to be contracted with herpes simplexI |

**6. Discussion**

The content of the speech were examined both in terms of quantity and quality at each assessed level. There was a clear difference in the amount of speech produced by the test-takers between the groups. In particular, a significant difference was observed between the low group and the other two groups. In terms of schematic structure, there was a noticeable relationship with the assessed level, and the higher groups produced a more complex structure than the low one. The findings of this study align with those of Brown et al. (2005). In their study, the amount of speech increased as the level went up, and the difference was largest between the low and the middle/high groups. The speech quality also had a clear relationship with proficiency levels. Simple structure was observed in the low level proficiency group and as the level increased, the speech data included more detailed information.

It is promising that the test was successful in differentiating between those candidates who need extra-support and those who do not. The speech data derived from the low group, member of which

are classified as requiring additional support in language and communication, contained a much smaller number of AS-units and clauses compared to that of the middle and high groups. This is not surprising, as the schematic structure of the low group was much less complex. The result implies that the speech data presented by the low group candidates lacks the amount of information required for a summary. The possible reasons for the result might be that 1) the test-takers struggle with organising the information provided by the input, or 2) the test-takers have difficulty in listening to and understanding the input stimuli. Given that the purpose of the test is diagnosis of the test-takers' language and communication skills, the cause should ideally be identified by the interviewer. It could be identified by an interviewer, asking probing questions to see whether test-takers are capable of producing more information, as was in fact suggested in an assessment guide for interviewers/raters in case of missing substantial content in students' summary. Examples were observed in the low group data where interviewer C prompted S3 and S4 for more details by asking questions. The rater noted, in the score sheets, that the content of S3 was 'too little'; while S4's was 'insufficient' and her 'comprehension was limited' which requires extensive language support. It is essential that such specific feedback is available for students at risk. In addition, it is important that all the raters have the same consensus of the interview process to diagnose students' weakness and supply sufficient information for their academic support.

The speech data of the high and middle groups did not reveal a significant difference in terms of amount. Both groups are considered to have produced a reasonable amount of information within their summaries; however, there was a noticeable difference in the quality of this information. The schematic structure derived from the high group was richer and more complex than that of the middle group. The test-taker assigned to the high group not only comprehended the details of the input, but was ready to present them. Connor & McCagg's (1983) findings support the results of this study. In their study, performances by 33 college students (22 non-native English speakers and 11 native English speakers) in a paraphrase task were

compared. Native speakers of English were found to pay more attention to detail, while the main points were perceived by both native and non-native speakers of English. Thus, it seems that the test-takers assigned to the high group in this study were closer to native level than those in the lower groups. Considering the outcomes of this analysis, tentative descriptors for judging the content of test-takers' oral summaries are presented in Figure 6.

One of the limitations of the analysis is that test-taker performances were compared on the basis that they were of the level assigned by the raters, which were derived from the existing scale lacking specific descriptors. In this regard, it is not clear if the scores assigned for the performances, and hence the proficiency level of candidates, can be trusted.

It is also possible that the way that the interviewer/raters related to the test-takers may have had an influence on the test-takers' performance as some language testing research suggests (e.g. Lazaraton (1996)), which was not discussed thoroughly in this paper. In most cases, the script examined in this study showed few interactions between interlocutors and test-takers due to the nature of the summarizing task. However, in the low group, as noted earlier, there were noticeable interactions between one interviewer/rater and test-takers, which intended to identify students' weakness. The effect of such interaction remains unclear and needs to be investigated further.

Another constraint of this study is that although significant differences were observed in the amount of speech produced at different proficiency levels, the accuracy of the content summary was not taken into account in the analysis. Micro-level distortions, such as giving a wrong time sequence ('four months later' instead of 'a few months later'), which had no major impact on overall meaning were occasionally observed in the summaries of the higher proficiency group. Macro-level distortions, such as misinterpretation of the issues of the case, were revealed in the summaries of the lower group. One possible reason for this outcome is that the summaries of the higher

group contained more utterances with detailed information, in consequence there were more chances of 'a slip of the tongue' when a candidate presented the summary. On the other hand, the distortion evident in summaries of the lower proficiency group is likely to be a result of candidates' difficulty with understanding the input stimuli. It is important that raters are given specific advice on how to deal with such distortions in allocating scores. More detailed study is needed to investigate such distortions, with a view to developing further rating descriptions with regard to test taker comprehension of input.

**Figure 6.** Possible Rating Scale Description.

| Content: Sufficiency and appropriateness of content | |
| --- | --- |
| Level 6-5 | sufficient amount of information, able to present key points as well as the details |
| Level 4 | appropriate amount of information, able to present most of the key points |
| Level 2-3 | insufficient amount of information, missing some key points, require help to organise their ideas |

## 7. Conclusion

This study explored test-taker performance on the HCST applying a range of discourse analyses. The content of the summaries, which has been an area of research interest in rating scale development for integrated tasks, was examined and provisional rating descriptors for the 'Content' criterion have been presented based on the empirical findings of this research.

The number of performance samples used in the current study is too small to allow us to generalise from these findings. However, it can be tentatively concluded that the current rating scale appears to be capable of distinguishing the students who need extra support in terms of their summarisation skills from those who do not. A

limitation of the study is the fact that it was conducted based on the assumption that scores assigned for the performance on which candidate proficiency levels were based, were both valid and reliable across the raters, however this cannot be guaranteed. This is particularly true for the content criterion, as the current rating descriptor contains no specific guidance for raters and may therefore have been used inconsistently.

This study has made a practical contribution to the HCST rating process. The study has also shown the usefulness of a close examination of test-taker performances in rating scale development for scale development purposes. The study needs to be extended in the interests of refining other criteria of the HCST including those of Task 2, the discussion section. Further analysis also needs to be undertaken with a larger set of discourse samples in order to validate the scores assigned using the refined rating scale. Such rating scale development can result in the HCST becoming a more valid, reliable and efficient diagnostic tool with potential benefits to all stakeholders.

## Acknowledgements

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.). *Language testing in the 1990s* (pp.71-86). London: Macmillan.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test taker performance on English for academic purposes speaking tests.* (Monograph Series MS-29). Princeton, NJ: Educational Testing Service.

Cohen, A. (1993). The role of instructions in testing summarizing ability. In D. Douglas & C. Chapelle (Eds.). *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium* (pp. 132-59). Washington DC: TESOL.

Connor, U., & McCagg, P. (1983). Cross-cultural differences and perceived quality in written paraphrases of English expository prose. *Applied Linguistics,* 4(3), 259-268.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing,* 11( 2), 125–44.

Eggins, S., & Slade, D. (1997). *Analysing casual conversation*. London: Cassell.

Educational Testing Service (2008). The TOEFL test - Test of English as a Foreign Language. Retrieved from http://www.ets.org

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354-375.

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

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13(2), 208-38.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Grove, E., & Brown, A. (2001). Tasks and criteria in a test of oral communication skills for first-year health science students: Where from? *Melbourne Papers in Language Testing*, 10(1), 37-47.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24-49.

Johns, A. M., & Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Applied Linguistics*, 11(3), 253-271.

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

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing*. 13(2), 151-172.

Lewkowicz, J. A. (1997). The integrated testing of second language. In C. Clapham & D. Corson (Eds.). *Encyclopedia of language and education 7* (pp. 121-130).

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

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

Miles M. & Huberman, M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks: Sage Publications.

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

Pollitt, A., & N. Murray (1996). What raters really pay attention to. In M.Milahovic & N. Saville (Eds.) *Performance testing, cognition, and assessment* (pp.74-91). Cambridge: University of Cambridge Local Examinations Syndicate/ Cambridge University Press.

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

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

Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers?. *Language Testing*, 24, 539-572.

**Appendix A:** Score sheet of the HCST

**Health Sciences Communication Skills Test**

Student Name_____ ID_____ School_____

Examiner_____        Date of interview_____

| A) LINGUISTIC PERFORMANCE | | *Score* | *Working notes* |
|---|---|---|---|
| **Language** | Range of structure & vocabulary; Breadth & precision of expression; Accuracy | ____ | |
| **Production** | Pronunciation; Intonation, Stress, Rhythm; Voice quality | ____ | |

**B) TASK ONE : PRESENTATION**

| | | | |
|---|---|---|---|
| **Organisation** | Macro structure of presentation | ____ | |
| **Content** | Sufficiency and appropriateness of content | ____ | |
| **Style** | Level of formality, tone; Nonverbal behaviour | ____ | |
| **Fluency & Coherence** | Sequencing, linking, clarity of ideas; Fluency of presentation | ____ | |
| **Comprehension of input** | | ____ | |

**C) TASK TWO : DISCUSSION**

| | | | |
|---|---|---|---|
| **Adequacy of participation** | Maintenance of interaction; Initiative, expansiveness | ____ | |
| **Quality of ideas** | Maturity or quality of ideas | ____ | |
| **Interpersonal skills** | Engagement, rapport; Nonverbal behaviour | ____ | |
| **Coherence & expression** | Clarity of ideas; Cohesion and coherence | ____ | |
| **Register & tone** | Level of formality; Politeness; Directness; Tone of voice | ____ | |

Diagnostic summary & Recommendations

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **Score codes** | Extremely weak | Major problems | Problems need attention; AT RISK | Minor problems | Fine | Excellent |