Investigating test-taker processes and task difficulty on an ESP reading test

Final report

Kellie Frost
Hyejeong Kim
John Pill
Catriona Fraser
Ute Knoch

Language Testing Research Centre
University of Melbourne
December 2012

Contents

Introduction

The Occupational English Test (OET) is a specific purpose test designed to evaluate the English-language competence of qualified medical and health professionals who wish to practise in an English-language context. It seeks to ensure that candidates are prepared, in language terms, for the world of work in their profession. A new summary cloze task has recently been included in the reading module of the Occupational English Test. The new task, which requires test-takers to skim and scan three to four short texts from a variety of sources, all dealing with a common topic, is designed to reflect the nature of current reading practices undertaken by health professionals in their work contexts, as identified in a prior job analysis by Elder, Harding and Knoch (2009).

The aim of the current project is to use verbal report methods to explore the construct validity of the new task by investigating if the processes test-takers engage in resemble those which the task is designed to elicit, as defined in the test specifications. The project provides insights into test-taker reading strategies specific to locating information across various texts, a type of second language reading which has thus far received little attention in the literature, as well as factors which impact on task difficulty.

Validity investigations such as the one outlined here should be based on practical argumentation theories, as set out by Kane (1992) and others (Bachman, 2005; Kane, Crooks & Cohen, 1999). An argument-based framework of test validation involves the development of an interpretive argument, whereby the way test scores are to be interpreted is explicitly stated, and a validity argument, or evaluation of the evidence in support of, and against, the assumptions inherent in the chain of inferences that lead from actual test performances to interpretations of test scores. According to such a framework, each link in the chain of inferences must be supported by evidence before it can be claimed that subsequent inferences, and any final interpretation, are valid. In terms of a broader validity argument, Xi (2008) and Chapelle, Enright and Jamieson (2010) describe six principal inferences that lie behind the interpretation of test results: domain description, evaluation, generalization, explanation, extrapolation and utilization.

The current project represents an investigation of the "explanation" inference. The first inference in the chain, "domain description", is the subject of a concurrent validity study (Macqueen et al., 2012), and the "evaluation" and "generalization" inferences, concerning the reliability of items and measures, are supported by rigorous and ongoing statistical analyses, conducted both pre-OET administration (post test trialling) and post-OET administration. The explanation inference, which is our focus here, is based on the expectation that scores are based on the reading construct that the task is designed to measure. Taking the framework set out by Chapelle, Enright and Jamieson (2010), the assumptions underlying this inference are that the reading knowledge, processes and strategies required to complete the new task are in accordance with theoretical expectations, as set out in the task specifications (see the methodology section, below), and that task difficulty is a systematic function of task characteristics.

Our project has the practical aims of defining more closely the theoretical construct underlying the type of professional reading that the new summary cloze task is intended to capture, and refining where necessary the design of the new task to ensure its appropriateness for providing evidence in support of conclusions about candidates' reading ability.

This final report begins with an overview of the relevant literature. Following the literature review, the specific research questions and methodology are described. Details of the results of the project, conclusions and recommendations are then reported.

Literature Review

As is now generally acknowledged, qualitative research which can shed light on test-taker knowledge, processes and strategies is required to supplement traditional score-based statistical analyses in order to build robust test validation arguments. Verbal reporting is an established methodology that serves this purpose, and has been widely used to investigate cognitive processes and strategies in relation to reading, as well as listening and writing performances. Gass and Mackey (2000) define verbal reporting as "gathering data by asking individuals to vocalise what is going through their minds as they are solving a problem or performing a task" (2000, p.13). In the domain of reading research, verbal reports have been widely used to examine first language (L1) reading behaviours (e.g., Cohen, 1986; Earthman, 1992; Fehrenbach, 1991; Gordon, 1990; Harmon, 2000), and a growing body of research has utilized the method in studies related to second-language (L2) reading comprehension. Some of the features of reading which have been investigated include: reading processes and strategy use (e.g., Anderson, 1991; Nevo, 1989; Cohen & Upton, 2007), cognitive processes (e.g., Lumley, 1993; Upton, 1997 & 1998), and the way in which reading strategies utilized in test situations differ from those employed in non-test contexts (e.g., Cordon & Day, 1996; Rupp et al., 2006).

The method is increasingly applied in testing-related research (see Lumley & Brown, 2005), and Green (1998) suggests that verbal reports may be used for a range of validation purposes. In a recent study, Cohen and Upton (2007) draw on verbal report methods to investigate the processes and reading behaviours of test-takers prompted by reading assessment tasks on the new TOEFL. In a similar vein, verbal reports will be used in the current project to investigate whether the hypothesised construct, articulated in the test specifications, is in fact being operationalized by the test task which candidates encounter. In other words, we aim to examine if the knowledge, processes and strategies elicited by the new reading task, as well as test-taker perceptions of task difficulty, are consistent with our theoretical expectations.

In light of this aim, a brief overview of the literature concerning reading comprehension processes and strategies is given below. Of further relevance is literature concerning the relationship between reading purposes and reading processes, which will follow. Some attention will also be given to Carver's (1990) "rauding" theory, which is based on the simple view of reading as a theoretical basis for understanding first language reading processes. According to the simple view of reading, reading comprehension is a product of the relationship between word recognition skills and listening comprehension abilities. Carver builds on this perspective, arguing that readers adjust their reading speed and processes depending on their reading purpose or goals. As the new OET reading task was designed to be completed within a limited time in an attempt to prompt reading behaviours of skimming and scanning, identified as key practices undertaken by health professionals in their work contexts, Carver's explanation of the relationship between reading rate and reading purpose is worthy of consideration here.

Reading comprehension processes and strategies

Broadly speaking, the literature on reading generally recognizes two levels of cognitive processing involved in reading comprehension: lower-level processing and higher-level processing. Lower-level processing, often referred to as bottom-up processing, includes word recognition, syntactic parsing and semantic-proposition encoding. Higher-level processing, often referred to as top-down processing, involves the ways in which readers make use of existing knowledge to interact with the text and to predict the text meaning, including directing attention to component skills (Alderson, 2000; Grabe, 2009). According to Grabe (2009), higher-level processing involves both the construction of what he calls a *text model of reader comprehension* and a *situation model of reader interpretation*, as well as *the executive control* of the reader, which is carried out as part of working memory and comprises goal setting, strategy use, and comprehension monitoring. The text model of reader comprehension "involves the combination of information from the currently formed proposition with the active meaning elements that have already been integrated into a network of ideas already activated from textual input" (Grabe, 2009, p. 40), whereas the situation model of reader interpretation involves combining background knowledge with the text, so it helps readers interpret a text in accordance with their own goals. The executive control "carries out key attentional processes and stores key information in the episodic buffer at the same time" (Grabe, 2009, p. 50). These attentional processes include (a) responding to reading goals and purposes, (b) applying strategies appropriately, (c) engaging metacognitive awareness and monitoring, (d) drawing on background knowledge as appropriate, and (e) supporting inferences for text processing and text evaluation (Grabe, 2009).

The relationship between reading purpose and reading processes

In general, broad models of reading such as that given above are based on reading for comprehension, or reading for learning in academic contexts (Grabe, 2009). Empirical research, however, has consistently shown that reading processes and strategies vary significantly depending on the purpose or goal of the reader (for example, van den Broek et al., 2001; Linderholm & van den Broek, 2002).

In an English L1 context, van den Broek et al. (2001) investigated if text recall and the generation of inferences by skilled readers varied depending on the reasons for reading. Using think-aloud protocols, they compared the number and types of inferences generated by college students (skilled readers) when reading for entertainment versus reading for study. They found that participants generated significantly more explanatory and predictive inferences, and more paraphrases and repetitions when reading for study. By contrast, reading for entertainment prompted significantly more evaluations of the text and "associations", defined by the authors as "retrieval of information not related to text coherence" (van den Broek et al., 2001, p. 1084). They also found that text recall was higher when participants were reading for study compared to reading for entertainment.

Van den Broek et al. (2001) explain their findings using the notion of "standards of coherence". According to this notion:

> "As readers proceed through a text, they maintain standards of coherence that act as criteria for comprehension. These standards or criteria, in turn, dictate the inferential activities in which the readers engage at each point during reading. The inferential activities that are employed directly determine the level of comprehension achieved" (p. 1082).

The authors suggest that standards might include referential, causal, global or thematic coherence, to name but a few possibilities. They further argue that these "standards of coherence" determine the ways in which lower-level and high-order reading processes are combined during reading in order to achieve a level of comprehension consistent with different reading goals.

In a study similar, Linderholm and van den Broek (2002) used verbal reports to compare inference generation in the same two reading conditions (study versus entertainment) between readers with low working memory capacity and those with high working memory capacity. Their findings were consistent with those reported by van den Broek et al. (2001) in that both participant groups demonstrated different patterns of inference generation depending on reading condition, although the low working memory capacity readers tended to rely on processes that placed the least demand on cognitive resources but that were not as efficient or effective in terms of comprehension outcomes. The authors expand on the abovementioned concept of "standards of coherence", arguing that both low- and high-working memory capacity readers adjust processing to meet these criteria, albeit with varying success. In terms of accepted models of reading, they further assert that "the strictness of the standard, in turn, affects the type and quality of mental representation that the reader constructs" (2002, p. 783).

Also using verbal reports, Horiba (2000) examined the effect on inference generation across different text (stories and essays) and task types (read freely and read for coherence) for L1 compared to L2 readers of Japanese, reporting results consistent with those of van den Broek et al., (2001) and Linderholm & van den Broek (2002), above. Both the L1 and L2 readers demonstrated different patterns of inference generation according to text type. For the essay, however, only the L1 group adjusted processing in response to task instructions. Horiba suggests that the L2 readers were less able to adapt to make processing more efficient on the more demanding task, although there was no difference between groups in terms of comprehension.

Grabe (2009, p. 8), based on a review of such studies, lists six major purposes for reading, each differing in terms of the level of detailed understanding (standards of coherence) likely to be deemed necessary by readers, and each therefore likely to involve different reading strategies and processes:

1. Reading to search for information (scanning and skimming)
2. Reading for quick understanding (skimming)
3. Reading to learn
4. Reading to integrate information
5. Reading to evaluate, critique and use information
6. Reading for general comprehension

Of particular relevance to the current study are purposes 1 and 2, above, both involving key search strategies of scanning, defined by Grabe as "identifying a specific graphic form", and skimming, or "building a simple quick understanding of the text" (2009, p 8). As Grabe notes, these search strategies were identified in earlier work by Guthrie and Kirsch (1987) and Guthrie (1988), detailed below. It is also expected that the new OET reading task will involve purposes 4 and 6, as test-takers are required to search for information and then to integrate the information into the summary text (purpose 4), and some of the more difficult items are intended to require test-takers to comprehend sections of source texts in order to be able to complete the paraphrasing in the summary text (purpose 6).

Guthrie and Kirsch (1987) investigated the notion that different reading purposes are likely to involve different reading processes and strategies in an L1 context of professional reading by electronics engineers and technicians. The authors compared tasks requiring readers to comprehend articles with tasks requiring readers to locate information in schematics, articles and work manuals. On the basis of factor analyses the authors concluded that reading processes involved in comprehension of articles were different to those involved in locating information. Further, they argue that while comprehension processes are specified in accepted models of L1 reading, locating information processes are not yet included in existing theoretical models and as a consequence are not represented by standardised tests of reading comprehension. In a subsequent paper, Guthrie (1988) proposed a cognitive processing model of reading to locate information in documents, which is relevant to the current study. The model involves stages of goal formation, category selection, information extraction, integration and recycling. Goal formation is the stage in which a question is formulated in relation to the information that is sought, category selection refers to the type of document likely to contain the required information, information extraction involves locating and taking note of the required information, integration refers to processes whereby the information is integrated with questions, and in the final stage, recycling, the reader decides if goals have been met and if not, returns to stage one to devise a new question. As will be seen below in the results and discussion and conclusion sections of this report, although this model is based on first language reading behaviours in a non- health professional context, it is also relevant to the reading behaviours that the new OET task is designed to elicit and can form the basis of a potential theoretical L2 model of reading to locate information.

Also of relevance to the current study is Carver's (1990) "rauding" theory. According to this theory, there are five qualitatively different processes relevant to reading in a first language depending on the reader's goal or purpose. As shown below in Table 1, these five processes involve different cognitive language components. Carver identifies "scanning" as the fastest process (or, in his terms, the fastest "gear"). He claims this process involves lexical accessing, whereby the reader locates a certain topic or target word. Carver found that effective scanning usually occurs at a rate of 600 words per minute (wpm) for L1 university students. The "skimming" process, the second fastest gear, entails lexical accessing and semantic encoding. Recognition of the word order and meaning in context is required for semantic encoding, and this occurs at an average rate of 450 wpm for L1 university students. So-called "rauding", the third gear, is the reading process engaged during ordinary or natural reading, and typically involves integrating sentences for understanding. The learning process, the second gear, takes more time (200 wpm on average) than the rauding process because thoughts or ideas must be first comprehended and then committed to memory. The slowest gear is a recall process in which individuals draw on and rehearse facts from memory. The typical rate for the recall process is 138 wpm.

Table 1. The five basic reading processes with goals, components, objective consequences, and typical rates for college students (Carver, 1990, p. 20).

Gear	Process	Goal	Components	Objective consequences	Rate (wpm)
5	Scanning (model)	Find target word	Lexical accessing	Correctly identify target word	600
4	Skimming (model)	Find anomalous words	Lexical accessing, semantic encoding	Correctly identify anomalous words in passage	450
3	Rauding	Understand the complete thoughts the writer intended to communicate	Lexical accessing, semantic encoding, sentence integrating	Correctly identify incomplete thoughts or anomalous sentences	300
2	Learning (model)	Know the information	Lexical accessing, semantic encoding, sentence integrating, idea remembering	Answer multiple choice questions on the passage	200
1	Recalling (model)	Recall the facts	Lexical accessing, semantic encoding, sentence integrating, idea remembering, fact rehearing	Write down exact words or facts from passage	138

Carver (1990) argues that "gear shifting" takes place depending on reading goals. When goals can be achieved without spending time on sentential integration, readers shift up to faster gears, such as skimming and scanning, while shifting down to slower gears, learning and recalling, when more time is required to accomplish their goals. In addition, gear shifting is influenced by the relative difficulty of reading materials. As would be expected, with more difficult texts, more time is needed for achieving goals. As noted above, Carver (1990) identifies rauding as natural reading. He makes the point that there is little variation between individuals or within individuals in terms of the rauding process, whereas other processes such as scanning, skimming, learning, and recalling, involve components which vary across and within individuals depending on conditions such as contexts, tasks, and materials.

In a study of L2 reading speed and comprehension, Haynes and Carr (1990) reported much lower levels of speed and comprehension for L2 readers compared to L1 readers of the same texts. They found that the mean reading speed was 86.5 wpm for adult Chinese L2 readers of English, with a comprehension average of 63.5% compared to the average L1 English readers' speed of 254 wpm, and comprehension average of 75.3%. Grabe (2009) also states that L2 students' reading speed in secondary university contexts may be between 80 and 120 wpm, which is one-half to one-third the rate of an L1 student. In a testing context, it is recognized that time constraints affect processing and therefore impact the cognitive validity of test tasks (Khalifa & Weir, 2009). Alderson (2000) suggests that in the context of language test development, tasks and task requirements should be designed according to which of the five processes identified by Carver (1990) that the test is intended to measure. It should also be noted that ordinary or 'natural' reading is a problematic notion in the context of second language use. It is unlikely that the sort of uniformity between and within individuals that Carver associates with "rauding" will occur in the case of adult second language readers (Koda, 2005).

In any case, given that different reading goals and purposes have been shown to be associated with different cognitive processes and strategies in L1 and to a limited extent in L2 contexts, the purpose associated with particular reading situations and tasks is clearly relevant to any definition of reading ability as a test construct. Furthermore, as limited investigations in L2 contexts so far suggest, the challenges faced by L2 readers, and the processes that they engage to meet their reading goals, are likely to be complex and varied compared to reading in a first language (Koda, 2005). In addition, previous investigations into L2 reading processes and assessment have, for the most part, focussed on reading strategy use on multiple choice comprehension tests (Phakiti, 2003 & 2008), with little attention paid to other task-types or to different reading goals in a testing context. The current study not only provides evidence in support of the validity of the new OET task as a measure of reading ability, but also offers valuable and novel insights into the processes involved in L2 reading to locate information across multiple texts, a type of reading found to be highly relevant to health professionals, and also likely to be of relevance to other professional and academic reading contexts.

Research Questions

In order to provide evidence of the construct validity of the new summary cloze reading task, this study addresses the following specific questions:

1. What reading processes, skills and strategies do items on the new task elicit from test takers at or above the pass level compared to test takers below the pass level on the OET reading sub-test?

2. What factors do test-takers perceive as affecting item difficulty?

Methodology

Participants

31 adult participants, for whom English is an L2, were recruited for this study with the assistance of the OET Centre. All participants have a health-related professional background, and were actual test-takers registered to take the OET at the administration which took place one week after the data were collected. Participant details are shown in Table 2, below.

Table 2. Participant details

No.	Gender	Profession	Years working in profession	Level*
1	F	Dentist	7 years	B
2	F	Doctor	18 months	B
3	F	Doctor	4 years	B
4	F	Nurse	4 years	C
5	F	Dentist	2 years	A
6	F	Nurse	4 years	B
7	M	Doctor	25 years	B
8	F	Nurse	3.5 years	C
9	F	Dentist	7 years	B
10	M	Doctor	3 years	B
11	F	Dentist	1 year	B
12	F	Pharmacist	3 years	B
13	M	Doctor	1 year	B
14	F	Pharmacist	3 years	B
15	F	Dentist	1 year	B
16	M	Doctor	3 years	B
17	F	Nurse	10 years	C
18	F	Doctor	5 years	B
19	F	Nurse	2 years	C
20	F	Doctor	8 years	B
21	F	Doctor	-	B
22	F	Nurse	6 years	C
23	M	Doctor	10 years	C
24	F	Doctor	5 years	B
25	F	Doctor	2 years	B
26	F	Dentist	4 years	B
27	F	Dentist	12 years	B
28	F	Nurse	9 years	B
29	M	Doctor	14 years	C
30	F	Nurse	4 months	D
31	F	Nurse	2.5 years	C

*Level = grade achieved on subsequent OET reading sub-test

Instruments and materials

The reading task

Participants were asked to complete a previously unseen retired version of the OET summary cloze reading task. The task included four short input texts from a variety of sources, each focussing on a different angle of the same topic, "Prison Health", and 29 test items (see Appendix A for the test version used in the study). Participants were required to skim and scan across the four texts in order to locate information required to complete gaps in the summary, and to then integrate the information into the summary text using up to three words per gap. Participants were asked to read and follow the same instructions that normally apply under live test conditions, and were allowed a strict time limit of 15 minutes.

Strategies questionnaire

When the task and verbal report session was complete, participants were asked to complete a questionnaire about their reading and test strategy use, adapted from questionnaires developed by Phakiti (2003 & 2008), see Appendix B.

Procedures

Verbal reports

A small pilot study was conducted to determine which of the following three possible approaches should be adopted for the collection of verbal reports: (i) participants complete the summary task in 15 minutes under test conditions, followed by stimulated recall; (ii) participants complete the task in three 5 minute chunks (15 minutes in total), with stimulated recall at the end of each 5 minutes; (iii) participants complete the task in six 2.5 minute chunks (15 minutes in total), stimulated recall at the end of each 2.5 minutes. Pilot study results showed that participants were unable to provide details beyond a post-test analysis of their strategy use after the full fifteen minutes. Even after five minute intervals, they were unable to recall their thoughts in any detail, whereas the shorter time chunks allowed greater memory access. Based on these results, the third option of dividing the task into six 2.5 minute intervals was adopted and is detailed below.

Participants completed the task in six timed 2.5 minute chunks (15 minutes in total). An immediate retrospective verbal report methodology was used at the end of each of the six timed sections. Participants were asked to "think aloud" about their reading behaviour and item responses in the preceding 2.5 minutes. The think aloud session was initiated with the prompt: *Tell me what you remember thinking during that time*, and followed up with: *Did you have any difficulty answering item(s) x, y etc/ what particular difficulty did you have* (see Appendix C for an outline of the protocol session). The texts and item responses of participants were used as stimulus materials. The verbal reports from each participant were audio recorded and transcribed.

Data analysis

For the purposes of analysis, participants were grouped according to whether they achieved a pass level (A or B) or not (C or D) on their subsequent sitting of the OET reading sub-test. The pass group included 22 participants, and the non-pass group included 9 participants.

Verbal report transcripts for each group were analyzed qualitatively in light of abilities listed in the task specifications, shown below in Figure 1, and the theoretical model of reading to locate information proposed by Guthrie (1988), discussed above. Questionnaire data were compiled and analysed quantitatively for each participant group.

Figure 1. Task specifications

The new reading task is designed to assess the following abilities:

- Locating specific information in a range of source texts
- Understanding the relationship between different types of information
- Understanding the conventions of different text types
- Identifying underlying concepts
- Drawing logical inferences
- Synthesizing information from different sources
- Differentiating main ideas from supporting information
- Identifying, distinguishing and comparing facts from a variety of text types
- Understanding the presentation of textual and numerical data
- Summarizing information for a non-medical audience
- Using contextual clues to determine text meaning and to supply missing information
- Recognizing paraphrase
- Using appropriate spelling and word forms

Results

Research question 1: What reading processes, skills and strategies do items on the new OET reading task elicit?

Analysis of verbal reports

On the basis of the verbal report data, we were able to identify a set of three main processes similar to those proposed by Guthrie (1988) that all participants across both groups tended to engage, albeit with varying levels of success, in order to complete the task items:

- **Locating information**
    - Predicting the type of information required (similar to Guthrie's "goal formation" stage)
    - Predicting the text likely to contain required information (similar to Guthrie's "category selection" stage)
    - Identifying salient features in summary text to direct scanning (similar to Guthrie's "information extraction" stage)
    - Locating these features in text ("information extraction" stage)
- **Integrating information** into summary (similar to Guthrie's "integration" stage)
- **Engaging test-taking strategies**

Many participants also engaged a fourth process of **verification** (similar to Guthrie's "recycling" stage) whereby they performed checks of summary text cohesion at both structural and semantic levels and amended structure or reformulated initial predictions if their responses did not appear to fit the summary meaning.

Verbal report data is organised below under headings referring to the first three main processes. Verification formed part of each of the other three processes, particularly for the pass group participants who seemed to readily access alternatives when predictions led to an initial incorrect response. In the cited extracts given throughout, "P" with a number refers to participant numbers (e.g., P1 = participant 1), and "R" refers to the researcher. As will be shown, participant verbal reports supported, for the most part, the list of abilities given in the task specifications.

Locating Information

All participants in both groups tended to focus on similar strategies for locating information. The strategy to identify salient features, such as numbers and names, or key words in the summary text, and then to scan for these in the source texts, was particularly prevalent. Most participants in both groups also read the headings of each of the source texts in order to identify the topic and predict where the required information was most likely located. The main difference between participants in the pass group and those in the non-pass group was that non-pass group participants tended to focus almost exclusively on scanning the source texts for numbers or names identified in the summary, and when such features were not available had difficulty identifying key words to search for, or were simply confounded as to how to proceed. The stronger participants tended to look beyond the gap in the summary and to skim the preceding and following summary content in order to better predict the sort of information required to ensure text cohesion, especially in the absence of salient features that corresponded directly with features in the source texts, and also in order to understand paraphrasing and identify possible synonyms in the source texts.

Pass group

As shown by the extracts below, pass group participants generally tried to predict the type of information that was required, usually by skimming the summary paragraph by paragraph and focussing on the words on either side of the gaps to find clues about topic, common collocations and word class. They tried to predict the text most likely to contain the information by skimming the text headings for topic, but also by attending to text genre and text conventions (as shown in the example from participant 18, below).

P1: I'm thinking about the statement that it says about the men who have been (xxx) four times and the other amount but I don't get it well so I needed to research a little bit for. And I read all the headlines of the four texts and it's about patient, patient health problems, (xxx) of environment and different kind of things.
R: So what did you do after you read the headlines?
P1: Ah read the first paragraph of the summary to know which, what kind of word I need to look for if it's a noun, or a verb, or a number and I always take a clue from the summary to find it in the text and it's my strategy.

P5: I just focused on one, you know, particular number. Yeah that's what I did and then I started to read but I didn't read the whole, just that portion where I could find that number. Um once I got this one, the next question was somewhere around here, it said 'four times', the next question was somewhere around here, it said 'four times' so… I was focusing on the summary, yep.
R: So you were picking some words like 'four times' out, is that what you did?
P5: Yeah. Four times and I found it here so… I tried to correlate that with here. Um then I read the sentence around, the whole sentence.

P6: The very first thing that I did was to check the text and look for some key words and then I approached the questions
R: So did you look at all four of the texts?
P6: Yeah. I just browsed the titles. Yeah. And then after that I just look for the first paragraph [in the summary] and after looking at the, the words, I search for the, which text will it be, will it be text one, text two or…

P12: Basically I was looking at the headlines of, of these, headlines about the prison and then I go to the summary, what is there talking about and there is a number, I just look for that number and then I just gradually followed the passage.

P18: Well first I look at all the articles, just only on titles, to know what it is about and what type of articles they are, if it is a research summary or if it is a table or graph or whatever.

As can be seen in the report from participant 10 below, pass group participants also engaged test strategies to manage time pressure. In this case, participant 10 reports directing attention towards items that allow for the strategy to scan quickly for highly salient features:

P10: Because it was the last minute so I have to run again to the numbers because numbers are the only thing which is, which pops out from the passage, it's easy to find the numbers. So again here it says '2004' so I have to find the passage again and I found 2004, 2004 here. So it's easier to find the numbers.

Non-pass group

As noted above, the non-pass group participants tended to focus almost exclusively on matching salient features such as numbers or key words directly from the summary text with features or words in the source texts. They also tended to look at text headings to identify topic, but didn't mention text types or genre as guiding their predictions:

P8: I read this summary and I, I important that there is “men in Australian prisons” and this is the number 25,240 so this summary is on part, text 1.

P17: Look at the, the, the figures and topics for every task, ah text, and the figures like percentages and numbers and things

P22: So I try to find clue, like number or something like capital letter because that’s easy to find it

P30: I scanned the words like ‘four times’ and first I scanned the word 25,240. I again go to questions, and look for key words and I found 1996 studies.

Integrating information

Once the required information was located, in order to “fit” information into the summary participants needed to either:

- copy the information directly into the summary
- make grammatical transformations
- use alternative lexical items (e.g., synonyms) or paraphrasing

As expected, items that allowed for direct copying were more likely to be answered correctly by participants in both groups than items that involved grammatical transformations or recognition of paraphrasing. The need to make transformations, access alternative lexical items or recognise paraphrasing was typically noticed by pass group participants, even if they were unable to fit information into the summary correctly, whereas non-pass group participants often did not notice that something more than simply copying words from source texts was required.

Pass group

When an alternative lexical item was required, high frequency synonyms were generally accessed by pass group participants, with less frequently encountered words causing more difficulty. When transformations or alternative lexical forms were required, pass group participants typically recognised that something needed to be modified and were generally able to do this successfully for common grammatical transformations (noun to adjective, for example) and for common synonyms:

P5: I tried to, you know, make it into a proper sentence by changing it, the percentage…
Yeah so it will be like approximately point one and point five so I just change it to five times. So yeah just trying to make a better proper sentence that sounds ok

P2: I found isolation, I just find the pairs of isolation and here frustrated, it’s very frustrated, I look for another adjective, I change to the adjective and write here.
R: Ok so which words did you change to the adjective?
P2: Ah angry, anxious
R: Oh I see, from anger to angry. Ok.

P6: I change it because the answer is ‘anxiety’ but it’s not, it’s supposed to be anxious and angry

P9: I have here ‘work’
R: So that’s number 15
P9: This word, it’s not in here, it’s not written ‘work’ but you can find ‘heaps of unemployment’ and they say ‘lack of opportunities for education’ for… so they don’t have ‘limited schooling’, schooling and rates of unemployment so you believe that it’s ‘work’. They have a lack of school, education and something else, you just believe that it’s ‘work’, it’s not written there.

It should be noted, however, that there were some exceptions. Participant 3, for example, does not recognise the need for a simple grammatical transformation and copied the words directly from the source texts:

P3: Actually 19 is the anger, anxiety.

Item 5, one of the items referred to by participant 6, below, required a complex grammatical transformation, from the noun 'release' to a passive verb construction 'are released from', or alternatively a synonymous active verb, 'leave', could have been used. Participant 6 did not recognise the need for a grammatical transformation, and copied 'release from' directly from the text. This was a particularly difficult item, however, and almost all pass group participants answered incorrectly. Only two participants answered correctly; most answered either 'release from' or 'released from':

R: Ok so you found one, two, four and five quite straight forward?
P6: Yeah because I search for the answers and it appears that the answers are vis-visible.
R: Did you change any of these?
P6: No
R: Ok so you took them straight from the text as they were.

Typically though, even if pass group participants were unable to complete the transformation or access suitable alternative lexical items, it was often clear that they had successfully located the correct information and recognised that a change was needed to preserve the source text meaning in the summary. For item 3, for example, participant 3 seemed to have located the required information, understood the text, and recognised that some sort of transformation was required to fit the information into the summary. Her answer, 'increased', is incorrect as it does not fit syntactically into the summary text (the answer should be 'higher' or 'greater' or 'larger' or 'bigger'), but she managed to capture the meaning:

P3: Mmm researched indicated that the (gap), what's the meaning, for which man, so I need to, first I, who had been four times so I try to find the four, present history was four times, four man, so I don't know the answer yet really.
R: Ok.
P3: Yeah so generally I concluded that that rate for man who has been in prison is four times increased than for men in general population. So finally I concluded.
R: So what made you conclude that answer?
P3: Yes I think so but I'm not sure. Trying to answer it.
R: Yep. What made you decide that that was the answer?
P3: Because, you know, the result of one major study showed that overall rate for men with prison history was four times that of a man in general community, so I can conclude that the men who live in prison history increase four times the rate of the general community of man, maybe Australia a bit different.

In relation to the same item, participant 5 also acknowledges that information from the source text needs to be changed to fit the summary, and was able to answer correctly:

P5: I'm sure that's not right. I need a lot of modification on that. Because sentence is not right, I mean, according to the research indicates that the [gap] men will have been in prison four times higher than, it's not the meaning of a proper sentence. I have to modify that.

Participant 12 is able to locate the information required for item 13, but fails to correctly transform the information grammatically to fit into the summary (she writes 'enter' instead of 'entering'), although she shows that she is aware of the need for a transformation:

P12: It says it is the poor health that is a concern for many of those prisoners. At first I look at text three, there is, I just look (poor/for) mental health. It goes to my eyes, poor mental health and I'm just scanning it and that is not poor health. Then I come back again to text one because it is related to this and I just sure that poor health so I just read and the poor health is already a concern for many of those, I just wrote entered prison. I just thinking prison, enter prison, the prisoner. So in summary we have to change sometimes grammar, noun, verbs.

For item 18, too difficult for most participants, the phrase 'long periods of isolation with little mental stimulus' from source text 3 was replaced in the summary with the phrase 'long periods of isolation [gap – item 18] much mental stimulus', so a word or words to negate 'much' were required to fit the summary and preserve the original meaning of 'with little' ('without' or 'with not', for example). Participant 15 recognises that a negative is required before 'much', but fails to recognise that 'with no much' is not grammatically correct:

P15: Then it comes to 18. That was also very easy because it said 'prisoners noted' so it is something related to the results. So I have to look for it in the results section so I focused my attention there and then it was something like mental or sorry, mental stimulus, I found the word and then yeah, I got it.

Also for item 18, participant 16 recognises that a modification is required to preserve the original meaning but is unable to access the word(s) needed so skips the item:

P16: it was saying that 'prisoners noted that periods of isolation blanks much mental stimulus' but I found that the answer must be in this one 'that long periods of isolation with little mental simulant, sorry stimulus' but I was unsure which words, like, which words should be like better to be with this much mental stimulus so I found it harder so that's why I wasn't sure which word I should put here. I'd like to like make my own word but I wasn't sure so that's why I just left it.

For the same item (18), participant 13 recognises that she needs to find a word to combine with 'much mental stimulus' that negates the meaning of 'much' and is one of only four participants who were able to answer correctly.

P13: And this one I (xxx) the isolation with not much, actually I just change it a little bit, it is little mental stimulus so 'not much'.

For items 24 and 25, although participant 5 answered correctly, she reported some difficulty integrating information, probably due to the fact that the paragraph in text 3 containing the information was paraphrased in the summary and needed close reading, and also that the several words from the text would potentially fit the gaps semantically and syntactically. In general, pass group participants tended to check regularly for text cohesion, and to recognise when a later part of the summary text made an earlier answer wrong:

P5: Yeah so yeah it was really tough. I couldn't focus and ah the answer which I thought was right-
R: this is for 24?
P5: Yeah. Spend with ah I was looking for, here, would be locked up for a long time and ensuing frustration would then be released on staff so they were lead to go out much so I thought it might be like 'spend with other inmates' but later to the paragraph when I was looking for this one I found that it is, it cannot be that but maybe with their family members, I don't know, so I just changed to the families and ah also less able to deal with bullying among prisoners.

For items requiring simple calculations based on figures presented in a table, many pass group participants expressed uncertainty or a reluctance to attempt the items due to the extra time they perceived would be required to perform the calculations. To correctly answer items 8, 9 and 10, for example, participants needed to look at two sets of figures presented in a table and make comparisons and many found this too demanding given the time constraints:

P3: In this I cannot manage very well because there's some calculating point in here so, you know, I need to find the words... On the other hand, I need to try and calculate so our brain cannot, you know, so many word or so many tasks it shocks me. We are under pressure, you know, only 15 minutes for as well and we need to answer 1 question within 30 seconds or something like that. So one minute or two minutes is very fast. If we cannot find the answer, you know, we can miss like the before answer we need to spend like two minutes for answer so it's nothing we cannot find and meanwhile our brain has to calculate how many, you know, how many rate or how many times but I think it's very difficult. I cannot do very well in this task.

Participant 18 also seemed to recognise what was required, but simply copied a figure directly from the table (0.5%) rather than performing the calculation:

P18: I knew what this was about but then I, I just, I was scared and I was supposed to figure out how much higher is prevalence of AIDS between people who are in prison and people who are, were, between the rest of people. So mathematics now.
R: So what did you put for your answer for that one?
P18: I first I, I just put the number
R: From the table?
P18: From the table yep. But then I realised that there is, in the sentence, I should write how much more. This is, that this is a occur in the people with history of imprisonment.
R: So what do you need to do now?
P18: I should calculate how much more is 0.5, how much more it is than 0.09.
R: Ok.
P18: Wow.

Finally, pass group participants also tended to recognise paraphrasing and, when they were unable to match it quickly with a corresponding section in the source texts, recognised that a more detailed understanding was required and that they needed to slow down and read sections of the source texts more closely. They often opted to skip these items due to the extra time burden, privileging a test-taking strategy (discussed further below) to try to maximise their test score by answering easier items first, or to guess (sometimes correctly) by inserting a word that they thought would fit syntactically and would make sense, even if they were unsure of whether or not they had captured the source text meaning:

P9: Now I have to stop and read because I can, I can't answer if I don't read and understand. It's pretty much, I know that it's here but I have to understand and put it here in the correct form. I might just have to change a verb for a noun or something like that but doesn't look like, looks like I have to understood, I have to understand and fill the gap with the correct word but I couldn't find it. But now I have to stop and read and understand. Stop, read and understand.
R: Right. And what, what, what gave you that idea, so what, why did you decide that you had to read it and understand?
P9: Because it's basically saying exactly the same thing here but in a different order or words. Maybe from the end to the beginning in this, you know, the opposite order and they say a few things here that I, I found here but doesn't fit properly so I have to understand with my own words.

Similarly, for item 3, participant 11 reports that she has located the information but recognises that the word(s) cannot be copied directly into the summary, and is unsure how to transform the information, so decides to skip the item rather than spend extra time trying to find an answer that fits the summary:

R: what was the difficulty with the third one?
P11: I couldn't find the answer in that line exactly but I thought instead of reading the same line three or four times it's better to go onto the next paragraph and find out the next answers.
R: Ok so you felt like you knew which line the answer was in?
P11: yeah

For item 22, participant 5 located the section of the source text containing the information needed, but failed to extract the words needed to fit the summary paraphrasing and made a guess (correctly) based on background knowledge of what would fit semantically and syntactically:

P5: 22, leading to sickness, yeah I was trying to find, I'm still not happy with the answer.
R: sickness?
P5: But still like that was, you know, the most accurate I could find at that moment so I just did that.
R: So why are you not happy with that answer?
P5: Because I didn't get time to read the whole sentence, it was just an imagination it might be that. I wasn't reading it rather I was thinking what would be the reason for absence, ok, due to sickness
R: Ok so you were guessing what you thought was more likely.
P5: Yeah rather than reading it because I didn't have time.

Non-pass group

As mentioned above, non-pass group participants often failed to recognise paraphrasing and overused the strategy of scanning the texts for salient items or key words from the summary. They often failed to identify the need for grammatical transformations, and relied too heavily on copying words directly from the source texts. Furthermore, they tended not to make checks to verify text cohesion.

For example, participant 8 does not identify any difficulty with items 13 and 14, but answers both incorrectly, inserting 'indigenous' instead of 'entering' for item 13 and 'health' instead of 'backgrounds' for item 14. The resulting sentence in the summary reads: – 'Poor health is already a concern for many of those **indigenous** prison because of issues connected with their **health**'. There is a lack of grammatical cohesion due to the first answer, and the repetition of 'health' negatively impacts text cohesion. In addition, the paraphrasing is inconsistent with the meaning of the source text.

Similarly, participant 19 did not identify any difficulty with items 1 through to items 9, but answered only one correctly (item 2). Problems were caused by failure to locate relevant information, a lack of word recognition and comprehension, and a failure to recognise grammatical inconsistencies:

R: Was it easy to find the answers or did you have any difficulty?
P19: No it was easy.
R: So can you tell me how you got your answer for question one? Where did you find it?
P19: I found from here 'Indigenous prisoners rarely represent total prison population' and according to the blank prison census means this Indigenous prison (census). Question three is prison is four times than, so here, prison history was four times that of the general community.

For items 18, 19 and 20, participant 23 demonstrated a failure to recognise a lack of cohesion and the need for grammatical transformations. In the summary sentence 'Prisoners noted that periods of isolation [gap – item 18] much mental stimulus left them feeling very frustrated, [gap – item 19] and [gap – item 20]', he inserted 'long periods' for the first gap (item 18), clearly copying words from the source text without realising that it didn't make sense, and failed to transform anger and anxiety into adjectives for items 18 and 19:

P23: So I found it here because I was in rush so I found it isolation ah ah long periods of isolation. So I put it same as and because I don't have exactly confidence can I say for a long period or not, I don't know. Anyway, that happened and after that isolation, and much mental stimulus and frustrated and I think it's here, yeah frustrated anger, feeling frustrated and anger. Until now it's easy.

As previously noted, many of the pass group participants also experienced difficulties with items requiring calculations. Typically, however, they recognised what was required because they had understood the summary text. For example, participant 18 from the pass group, cited above in relation to item 8, understood that a comparison was being made in the summary paraphrasing even though she failed to make the required calculation. For the same item, participant 29 in the non-pass group, as with most participants in this group, did not express uncertainty or difficulty but did not seem to understand what was required, answering 0.4 because it is lower than 0.5 (the correct answer was '5 times': 'The rate of AIDS among prisoners was over **5 times** the rate for the population as a whole'):

P29: This is correct this is 0.4 because it's right over, it's not over this is 0.5, and because that I write 0.4

Test-taking strategy

The main test-taking strategy used by all participants was to skip difficult items in order to have time to answer easier items and thereby maximise their test score. As noted previously, time pressure led most participants to use a strategy whereby they filled the "easy" gaps – those which allowed information to be copied directly – and skipped over gaps requiring transformation or involving paraphrasing when this could not be achieved almost automatically. The reading purpose of "finding information" is clearly artificial in the sense that the real goal is obviously to correctly fill as many gaps as possible:

P5: I was trying to find the answer for 26 which I couldn't on a glance. Because I realised that I, I kept on reading so I realised the (cost) is going to be easier and that will give me two or three more blanks rather than going for just this one which I have no idea still.

P4: I wasn't sure about what is the correct word to fit in that gap so yeah I just skip it. I just thought maybe good idea to just move on to the other questions so I can answer as much as possible.

R: I see you've jumped over question 3 there.
P6: Yeah because I just don't want to waste time so if I ever find a blank difficult, question difficult, I'm trying to answer the other questions first, so yeah.

P10: I couldn't' find the answer for four and five so I just skipped that passage and went to the next one. I didn't want to waste time sitting and finding answers for that one.

P11: I don't want to waste more time searching for the singular answer rather than like, I go to the next one so that I finish off the easiest lines then I can go back to that first one if I have time. Instead of wasting time and focusing on answer, I feel it's better to go on.

P14: Incidentally I thought I should now move to the next paragraph, next summary because I'm not getting the answer so maybe I should just skip this and I should start focusing the next.

Though participants sometimes noted that they intended to return to the items they had skipped when they reached the end of the test, most of them were not able to do so within the 15 minutes allowed.

Analysis of strategies questionnaire responses

Participant responses on how frequently they used strategies were canvassed via a five-point Likert scale (from 1 *Never* to 5 *Always*) as shown in Table 3 below. Overall, the participants' responses from the two groups show a similar tendency in terms of the use of each strategy which seems to reflect the characteristics of the summary task, and their responses were consistent with data collected via verbal reports.

More frequently used strategies

Responses reported as Likert scale 3.5 and above form high frequency use. The participants reported having planned how to complete the test and followed their plan before starting the test (Q1, M=3.6). A stronger tendency was reported towards more frequent use of scanning (Q2, M=4.0). They reported endeavouring to understand the relationships between ideas in the texts and task (Q6, M=3.5), and a considerably stronger tendency was reported in terms of using the titles of the texts, tables, or figures to aid comprehension (Q7, M=4.7) with no responses of *Never* or *Sometimes* and *Always* from the majority. Scanning and skimming topics and main ideas in each text separately was also reported as a frequently used strategy (Q13, M=4.3), and many participants reported having tried to identify easy and difficult test items (Q17, M=3.5). They also stated that they were aware of what strategies they were using in the test (Q18, M=3.6), how much of the test remained to be completed (Q22, M=3.7) and the need to plan a course of action (Q24, M=3.6). They reported that they knew when they should read more quickly or carefully (Q25, M=4.2) with no response of *Never* to this question.

Strategies used with medium frequency

Responses reported as between Likert scale 2.6 and 3.4 were considered medium frequency use. Responses to the question about whether they had summarised the main information in the texts in their head were relatively equally distributed across scale items (Q4, M=3.1), and a marginally lesser use of note taking strategies was reported (Q5, M=2.8). The participants tended to use their knowledge of information structure in English (Q8, M=3.4), to predict what was going to come next during reading (Q9, M=2.8), to analyse author's meaning and intention (Q10, M=3.1), and to try to understand the texts and tasks regardless of their vocabulary knowledge (Q12, M=3.3) – all at the level of medium frequency. Similarly, they reported having tried to find topics and main ideas by scanning and skimming across all the texts at the same time (Q14, M=3.2), and they reported using strategies of checking their own performance and progress during the text (Q19, M=3.3), and guessing meanings of unknown words (Q20, M=3.1). Lastly, they stated that they knew which information was more or less important (Q23, M=3.3).

Less frequently used strategies

Responses reported as Likert scale 2.5 and below were classified as low frequency use. Translating the texts and task into their first language was a less frequently used strategy (Q3, M=1.8) and none of the participants reported having *Always* translated. Marginally more frequent use of strategies was reported with regard to spending more time on difficult items (Q11, M=2.4), reading through the texts and task several times to better understand them (Q15, M=2.4), and relating the information from the texts or task to their prior knowledge or experience (Q16, M=2.1) with no response of *Always* to the last question. Taking time to review their responses to the items is also included in less frequently used strategies (Q21, M=2.5).

Table 3. Participants' responses to strategies questionnaire by group

Your thinking	Responses	Levels A & B (n=22)	Levels C & D (n=9)	Total (N=31)
Q1. When I started the test, I planned how to complete it and followed my plan.	never	2	0	2
	sometimes	5	2	7
	often	2	1	3
	usually	6	2	8
	always	7	4	11
Q2. I scanned through the reading test before I actually started to complete it.	never	1	0	1
	sometimes	3	0	3
	often	3	3	6
	usually	5	1	6
	always	10	5	15
Q3. I translated the reading texts and task into my first language.	never	12	4	16
	sometimes	6	2	8
	often	0	1	1
	usually	3	1	4
	always	–	–	–
Q4. I summarised in my head the main information in the texts.	never	2	1	3
	sometimes	7	1	8
	often	3	3	6
	usually	8	1	9
	always	1	3	4
Q5. I made short notes or underlined main ideas during the test.	never	7	2	9
	sometimes	6	2	8
	often	0	2	2
	usually	5	0	5
	always	4	3	7
Q6. I tried to understand the relationships between ideas in the texts and task.	never	0	1	1
	sometimes	6	3	9
	often	2	1	3
	usually	8	2	10
	always	6	2	8
Q7. I used the titles of the texts, tables or figures to help comprehend them.	never	–	–	–
	sometimes	–	–	–
	often	1	1	2
	usually	4	2	6
	always	17	6	23
Q8. I used my knowledge of how information is structured in English to comprehend the texts.	never	3	1	4
	sometimes	3	0	3
	often	2	2	4
	usually	10	4	14
	always	3	2	5
Q9. I predicted what was going to come next while I was reading the texts.	never	2	2	4
	sometimes	10	2	12
	often	3	1	4
	usually	6	2	8

	always	1	2	3
Q10. I analysed what the author meant or tried to say in the texts.	never	4	2	6
	sometimes	4	2	6
	often	4	1	5
	usually	4	1	5
	always	6	2	8
Q11. I spent more time on difficult items.	never	8	2	10
	sometimes	6	3	9
	often	2	1	3
	usually	3	1	4
	always	2	2	4
Q12. I tried to understand the texts and task regardless of my vocabulary knowledge.	never	2	1	3
	sometimes	5	3	8
	often	3	0	3
	usually	6	4	10
	always	6	1	7
Q13. I tried to find topics and main ideas by scanning and skimming each text separately, one after the other.	never	1	0	1
	sometimes	1	1	2
	often	1	1	2
	usually	7	2	9
	always	12	5	17
Q14. I tried to find topics and main ideas by scanning and skimming across all the texts at the same time.	never	3	2	5
	sometimes	3	2	5
	often	4	1	5
	usually	5	3	8
	always	6	1	7
Q15. I read through the texts and task several times to better understand them.	never	6	3	9
	sometimes	9	1	10
	often	1	2	3
	usually	5	2	7
	always	0	1	1
Q16. I related the information from the texts or task to my prior knowledge or experience.	never	11	1	12
	sometimes	5	1	6
	often	4	4	8
	usually	2	1	3
	always	–	–	–
Q17. I tried to identify easy and difficult test items.	never	4	0	4
	sometimes	4	0	4
	often	3	2	5
	usually	7	1	8
	always	4	6	10
Q18. I was aware of what strategies I was using in the test.	never	2	0	2
	sometimes	5	0	5
	often	4	1	5
	usually	8	3	11
	always	3	5	8
Q19. I checked my own performance and progress while completing the test.	never	2	0	2
	sometimes	7	2	9
	often	3	2	5

	usually	6	2	8
	always	4	3	7
Q20. I guessed meanings of unknown words.	never	3	1	4
	sometimes	6	1	7
	often	3	1	4
	usually	10	5	15
	always	0	1	1
Q21. I took time to review my responses to the items	never	5	0	5
	sometimes	13	3	16
	often	0	2	2
	usually	2	3	5
	always	2	1	3
Q22. I was aware of how much the test remained to be completed.	never	3	0	3
	sometimes	5	0	5
	often	3	1	4
	usually	3	3	6
	always	8	5	13
Q23. I knew which information was more or less important.	never	0	2	2
	sometimes	8	1	9
	often	3	1	4
	usually	5	2	7
	always	5	3	8
Q24. I was aware of the need to plan a course of action.	never	1	0	1
	sometimes	5	0	5
	often	4	2	6
	usually	10	2	12
	always	2	5	7
Q25. I knew when I should read more quickly or carefully.	never	–	–	–
	sometimes	3	1	4
	often	2	0	2
	usually	8	2	10
	always	9	6	15

Research question 2: What factors do test-takers perceive as affecting item difficulty?

As is clear from the extracts presented earlier in this section, items requiring complex grammatical transformations, alternative lexical items, or involving paraphrasing of source texts were perceived by all participants as more difficult than items where words could be directly copied from source texts.

Many participants also commented on time pressure as a source of difficulty, especially when information required in the summary was difficult to locate in the source texts. Information appeared to be difficult and time consuming to locate when there were no key words or salient features in the source texts that corresponded directly with words or salient features in the summary, as was the case when heavy paraphrasing was used. For example, a lack of salient features made items 13-16 difficult for almost all participants, regardless of group. The sentences in the texts containing the required information were paraphrased in the summary, which hindered the common strategy to scan for key words from the summary in the texts. Further, the words "health" and "prison" in the sentence surrounding gap 13 in the summary initially led all participants to text 3, as the same words featured in the title of the third text. Most participants gave up when the information could not be found in text 3. Some realised after a slow search of text 3 that the information might be found in text 1, but were unable to locate the information by quickly skimming, and given the time pressure decided to skip the items rather than to engage in the slower, more detailed reading which was required even when the correct text was identified.

Some also commented that the time constraints and the task instructions to skim rather than to read the entire texts were not always appropriate to task demands:

P25: I don't know how to do scanning and skimming technique they say to follow but to follow scanning and skimming takes- you have to look one word and you have to go to text and you have to find answer straight away but in here I don't find answer straight away. I have to take out the meaning for some answer I have to take out the meaning and write in my own words and just scanning and skimming I try to look at the answers but that did not help me in this text because there is no directly. Ah directly you can't scan and skim and find like that, it's not in text

P26: I think the problems for part A because I realise not only the reading of text it needs actually to get a good comprehension of text but I think sometimes it's really hard in 15 minutes just to get a good comprehension of the text so if I could find the exact words it's easier. So I try to find the words or the synonym words but sometimes I have to make up just some questions so at that time I have to read carefully and then find the exact words.

As mentioned, many participants also identified difficulties with having to perform simple calculations under time pressure. In many cases it was clear that participants had understood the source and summary texts, and knew what was required, but had either skipped the items or answered incorrectly because they baulked at the mental arithmetic involved.

Conclusions

The primary aim of the current study was to explore the construct validity of the new OET summary cloze reading task by investigating if the processes test-takers reported engaging in resembled those which the task is designed to elicit. Data collected using verbal reports and a strategies questionnaire provide evidence in support of the validity of the task, as many of the reading processes and strategies reported by test-takers mirrored theoretical expectations, and called upon abilities listed in the task specifications. As anticipated, Guthrie's (1988) model of reading to locate information was amenable to adaptation to capture the processes involved in completing our task. Most participants reported processes that corresponded to Guthrie's stages of "goal formation", "category selection" and "information extraction" (labelled here as "locating information") as well as "integration" (labelled here as "integrating information"). These macro processes could be readily associated with the use of most of the abilities listed in the existing task specifications, and offer a means of organising the existing taxonomy of skills around a theoretically robust processing model, and of further specifying task design and item difficulty components. As expected, stronger participants also reported verifying text cohesion as they progressed through the task, and appeared to access a greater range of strategies with greater flexibility than the non-pass group participants who, as noted, tended to over rely on a scan and copy strategy. Findings also supported van den Broek et al.'s (2001) notion of "standards of coherence", as many of the pass group participants reported varying reading speed and purpose depending on the level of comprehension deemed necessary to ensure summary text coherence and consistency with meaning in the source texts. Consistent with existing second language reading literature (Grabe, 2009; Koda, 2005), our findings also indicate a lack of uniformity among pass group participants, with the cause of common difficulties (difficulties in locating information and integrating information) varying from participant to participant due to gaps in linguistic knowledge and resources. Non-pass group participants generally displayed a lack of word recognition and syntactic knowledge in relation to the stronger group, which provides further evidence of the validity of the task. There was also evidence to suggest that items were skipped due to difficulty as well as time pressure, thereby offering some support for the speeded nature of the task, although, as expected, participants across both groups varied greatly in terms of reading and item completion speeds, with many of the pass group participants completing no more than two-thirds of the items.

In relation to the speeded nature of the task, our results suggest that time pressure had some impact on test-taker behaviour, leading participants who appeared to have successfully comprehended the text and summary information to skip items that required time consuming cognitive processes. Such a finding supports claims made by Khalifa and Weir (2009), who suggest that time constraints affect processing and consequently the validity of test tasks; this may therefore indicate a threat to the validity of the summary cloze task. There was also evidence from most participants to indicate that items requiring mathematical calculations were avoided because of non-reading related difficulties (although the task specifications do refer to "understanding the presentation of … numerical data"). Even many of the pass group participants who had seemingly understood the texts and item requirements either skipped or answered these items incorrectly. These potential threats to validity warrant further quantitative and qualitative investigations to determine if such effects are significant, possibly requiring a modification to the marking procedures, which currently do not distinguish between incorrect and omitted (no response) items, and more detailed item writing specifications.

Finally, in terms of a validity argument as specified by Kane and associates (1992, 1999), Bachman (2005) and more recently by Xi (2008) and Chapelle, Enright and Jamieson (2010), the current study provides evidence in support of our inference that the reading knowledge, processes and strategies required to complete the summary cloze task are consistent with theoretical expectations and with the skills and abilities set out in the test specifications. Qualitative differences in reading processes and strategies were noted between participants in the pass group compared to those in the non-pass group. As discussed above, however, findings also suggest that task difficulty, while predominately a systematic function of task characteristics, might also be determined in part by construct-irrelevant factors such as time pressure and ability to perform mathematical calculations. As discussed above, these factors warrant further investigation.

Recommendations

On a practical level, we aimed to use insights gained from the study to inform refinements to the task specifications and to thereby enhance future task design. Based on our findings, we suggest the development of more detailed task specifications, including definitions of item types based on the processes associated with locating and integrating information so that a measured range of item difficulty can be included in each task version (for example, easy items = salient features + direct copy; medium difficulty = simple grammatical transformations, high frequency synonyms; most difficult = understanding complex paraphrasing).

As noted above, it is also possible that items requiring simple mathematical calculations represent a source of construct irrelevant variance and their inclusion should perhaps be minimised until further investigations are conducted. We also suggest further investigations to determine if a lack of response is primarily a product of item difficulty or of time constraints.

References

Alderson, C. J. (1984). Reading in a foreign language: A reading problem or a language problem? In C. J. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language*. London: Longman.

Alderson, C. J. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal, 75*(4), 460-472.

Bachman, L.F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1-34.

Carver, R. (1990). *Reading rate: A review of research and theory*. San Diego, CA: Academic Press, Inc.

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

Cohen, A. D. (1986). Mentalistic measures in reading strategy research: Some recent findings. *English for Specific Purposes*, 5, 131-145.

Cohen, A. D. (1996). Verbal reports as a source of insights into second language learner strategies. *Applied Language Learning, 7*(1&2), 5-24.

Cohen, A. D. (1998). Strategies and processes in test taking and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Cohen, A. D., & Upton, T. A. (2007). "I want to go back to the text": Response strategies on the reading subtest of the new TOEFL. *Language Testing*, 24, 209-250.

Cordon, L. A., & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology*, 88, 288-295.

Earthman, E. A. (1992). Creating the virtual work: Readers' processes in understanding literary texts. *Research in the Teaching of English*, 26, 351-384.

Elder, C., Harding, L., & Knoch, U. (2009). *OET reading revision study: Final report*. Unpublished report, Language Testing Research Centre, University of Melbourne.

Fehrenbach, C. R. (1991). Gifted/average readers: Do they use the same reading strategies? *Gifted Child Quarterly*, 35, 125-127.

Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gordon, C. J. (1990). Modeling an expository text structure strategy in think alouds. *Reading Horizons*, 31, 149-167.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

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

Guthrie, J.T. (1988). Locating information in documents: Examination of a cognitive model. *Reading Research Quarterly*, 23, 178-199.

Guthrie, J.T., & Kirsch, I.S. (1987). Distinctions between reading comprehension and locating information in text. *Journal of Educational Psychology*, 79, 220-227.

Harmon, J. M. (2000). Assessing and supporting independent word learning strategies of middle school students. *Journal of Adolescent and Adult Literacy*, 43, 518-527.

Haynes, M., & Carr, T. H. (1990). Writing system background and second language reading: A component skills analysis of English reading by native speakers of Chinese. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 375-421). New York, NY: Academic Press.

Horiba, Y. (2000). Reader control in reading: Effects of language competence, text type and task. *Discourse Processes*, 29, 223-267.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.

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

Khalifa, H., & Weir, C. (2009). *Examing reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge: Cambridge University Press.

Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology*, 94, 778-784.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, 10, 211-234.

Lumley, T., & Brown, A. (2005). Research methods in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 833-856). Mahwah, NJ: Lawrence Erlbaum Associates.

Macqueen, S. , Yahalom, S., Kim, H., & Knoch, U. (2012). *Exploring writing demands in healthcare settings*. Unpublished report, Language Testing Research Centre, University of Melbourne.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing*, 6, 199-215.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, 20, 26-56.

Phakiti, A. (2008). Strategic competence as a fourth-order factor model: A structural equation modelling approach, *Language Assessment Quarterly*, 5, 20-42.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing*, 23, 441-474.

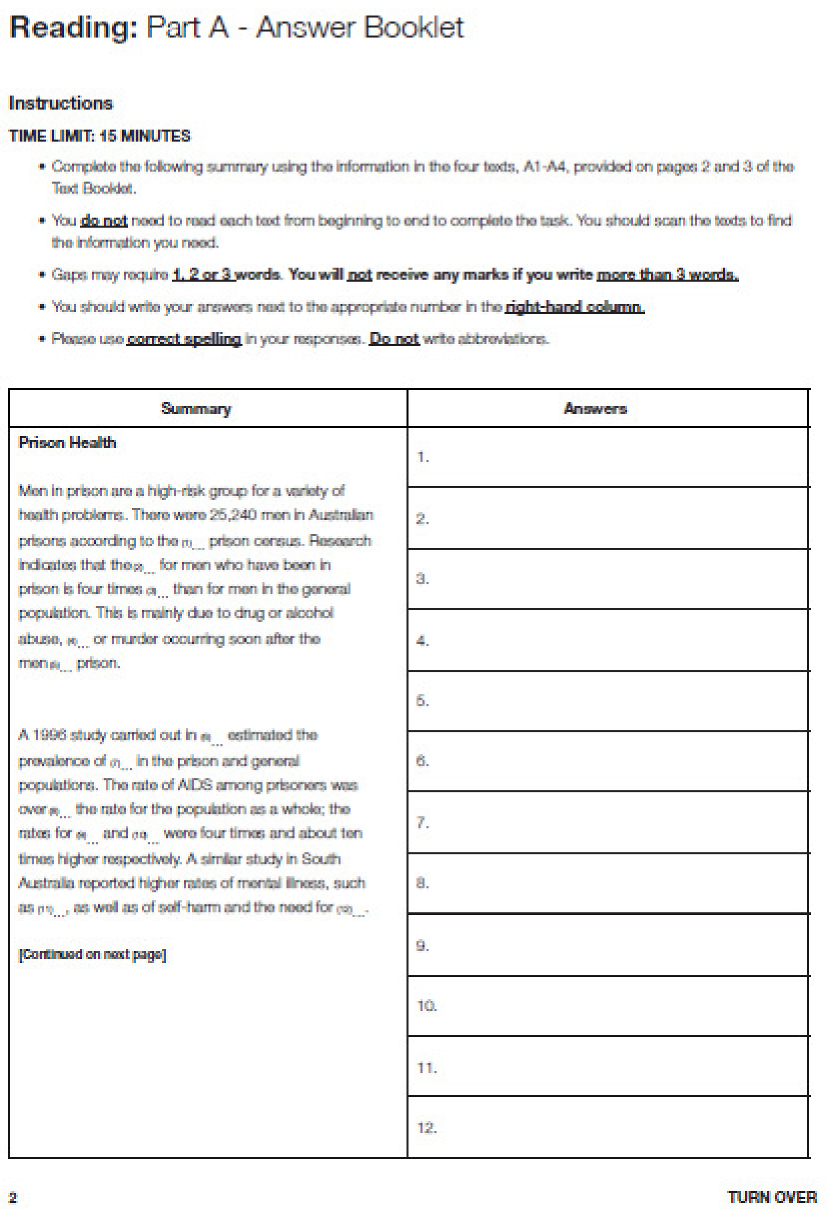
Upton, T. A. (1997). First and second language use in reading comprehension strategies of Japanese ESL students. *TESL-EJ*, 3, A-3.

Upton, T. A. (1998). "Yuk, the skin of insects!" Tracking sources of error in second language reading comprehension. *Journal of College Reading and Learning*, 29, 5-20.

van den Broek, P., Lorch, R. F. Jr., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition*, 29, 1081-1087.

Xi, X. (2008). Methods of Test Validation. In E. Shohamy and N.H. Hornberger (eds). *Encyclopedia of Language and Education*, 2nd edition, Volume 7: Language Testing and Assessment, pp. 177-196.

Appendix A – the task



Reading: Part A - Answer Booklet

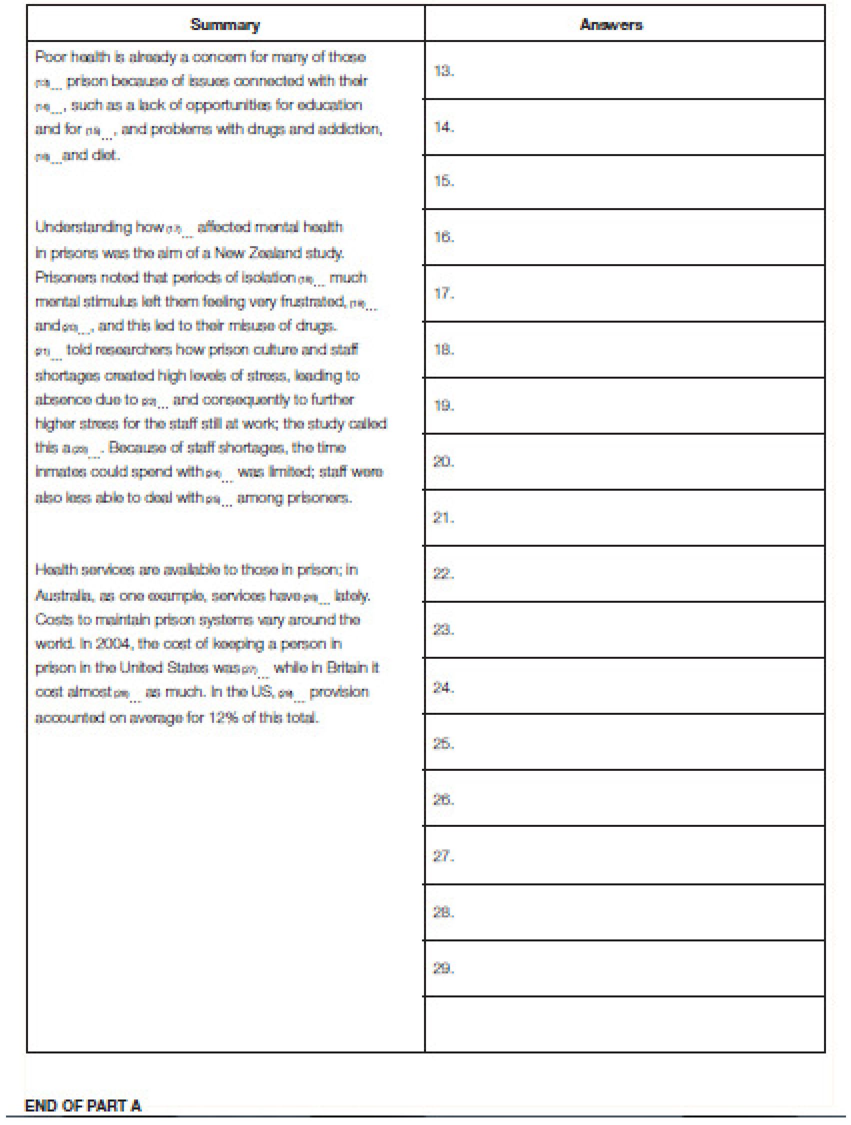
Instructions

TIME LIMIT: 15 MINUTES

- Complete the following summary using the information in the four texts, A1-A4, provided on pages 2 and 3 of the Text Booklet.
- You do not need to read each text from beginning to end to complete the task. You should scan the texts to find the information you need.
- Gaps may require **1, 2 or 3 words. You will not receive any marks if you write more than 3 words.**
- You should write your answers next to the appropriate number in the **right-hand column.**
- Please use **correct spelling** in your responses. **Do not** write abbreviations.

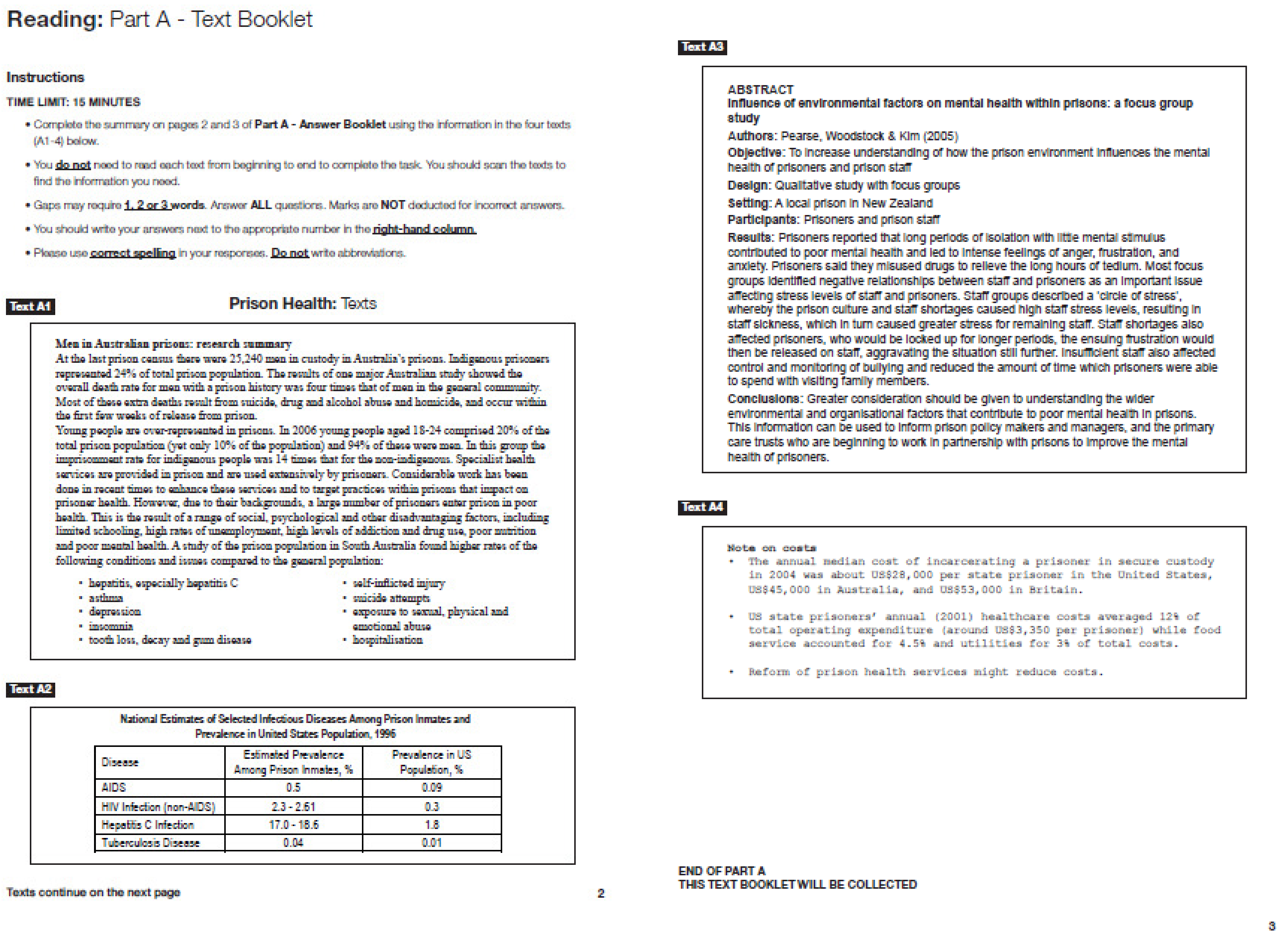
Summary	Answers
Prison Health Men in prison are a high-risk group for a variety of health problems. There were 25,240 men in Australian prisons according to the (1)... prison census. Research indicates that the (2)... for men who have been in prison is four times (3)... than for men in the general population. This is mainly due to drug or alcohol abuse, (4)... or murder occurring soon after the men (5)... prison. A 1996 study carried out in (6)... estimated the prevalence of (7)... in the prison and general populations. The rate of AIDS among prisoners was over (8)... the rate for the population as a whole; the rates for (9)... and (10)... were four times and about ten times higher respectively. A similar study in South Australia reported higher rates of mental illness, such as (11)..., as well as of self-harm and the need for (12).... **[Continued on next page]**	1.
	2.
	3.
	4.
	5.
	6.
	7.
	8.
	9.
	10.
	11.
	12.

2 TURN OVER



Summary	Answers
Poor health is already a concern for many of those (13)... prison because of issues connected with their (14)..., such as a lack of opportunities for education and for (15)..., and problems with drugs and addiction, (16)... and diet. Understanding how (17)... affected mental health in prisons was the aim of a New Zealand study. Prisoners noted that periods of isolation (18)... much mental stimulus left them feeling very frustrated, (19)... and (20)..., and this led to their misuse of drugs. (21)... told researchers how prison culture and staff shortages created high levels of stress, leading to absence due to (22)... and consequently to further higher stress for the staff still at work; the study called this a (23)... . Because of staff shortages, the time inmates could spend with (24)... was limited; staff were also less able to deal with (25)... among prisoners. Health services are available to those in prison; in Australia, as one example, services have (26)... lately. Costs to maintain prison systems vary around the world. In 2004, the cost of keeping a person in prison in the United States was (27)... while in Britain it cost almost (28)... as much. In the US, (29)... provision accounted on average for 12% of this total.	13.
	14.
	15.
	16.
	17.
	18.
	19.
	20.
	21.
	22.
	23.
	24.
	25.
	26.
	27.
	28.
	29.

END OF PART A



Reading: Part A - Text Booklet

Instructions

TIME LIMIT: 15 MINUTES

- Complete the summary on pages 2 and 3 of **Part A - Answer Booklet** using the information in the four texts (A1-4) below.
- You do not need to read each text from beginning to end to complete the task. You should scan the texts to find the information you need.
- Gaps may require **1, 2 or 3 words**. Answer **ALL** questions. Marks are **NOT** deducted for incorrect answers.
- You should write your answers next to the appropriate number in the **right-hand column.**
- Please use **correct spelling** in your responses. **Do not** write abbreviations.

Text A1

Prison Health: Texts

Men in Australian prisons: research summary

At the last prison census there were 25,240 men in custody in Australia's prisons. Indigenous prisoners represented 24% of total prison population. The results of one major Australian study showed the overall death rate for men with a prison history was four times that of men in the general community. Most of these extra deaths result from suicide, drug and alcohol abuse and homicide, and occur within the first few weeks of release from prison.

Young people are over-represented in prisons. In 2006 young people aged 18-24 comprised 20% of the total prison population (yet only 10% of the population) and 94% of these were men. In this group the imprisonment rate for indigenous people was 14 times that for the non-indigenous. Specialist health services are provided in prison and are used extensively by prisoners. Considerable work has been done in recent times to enhance these services and to target practices within prisons that impact on prisoner health. However, due to their backgrounds, a large number of prisoners enter prison in poor health. This is the result of a range of social, psychological and other disadvantaging factors, including limited schooling, high rates of unemployment, high levels of addiction and drug use, poor nutrition and poor mental health. A study of the prison population in South Australia found higher rates of the following conditions and issues compared to the general population:

- hepatitis, especially hepatitis C
- asthma
- depression
- insomnia
- tooth loss, decay and gum disease
- self-inflicted injury
- suicide attempts
- exposure to sexual, physical and emotional abuse
- hospitalisation

Text A2

National Estimates of Selected Infectious Diseases Among Prison Inmates and Prevalence in United States Population, 1996

Disease	Estimated Prevalence Among Prison Inmates, %	Prevalence in US Population, %
AIDS	0.5	0.09
HIV Infection (non-AIDS)	2.3 - 2.61	0.3
Hepatitis C Infection	17.0 - 18.6	1.8
Tuberculosis Disease	0.04	0.01

Texts continue on the next page

Text A3

ABSTRACT

Influence of environmental factors on mental health within prisons: a focus group study

Authors: Pearse, Woodstock & Kim (2005)

Objective: To increase understanding of how the prison environment influences the mental health of prisoners and prison staff

Design: Qualitative study with focus groups

Setting: A local prison in New Zealand

Participants: Prisoners and prison staff

Results: Prisoners reported that long periods of isolation with little mental stimulus contributed to poor mental health and led to intense feelings of anger, frustration, and anxiety. Prisoners said they misused drugs to relieve the long hours of tedium. Most focus groups identified negative relationships between staff and prisoners as an important issue affecting stress levels of staff and prisoners. Staff groups described a 'circle of stress', whereby the prison culture and staff shortages caused high staff stress levels, resulting in staff sickness, which in turn caused greater stress for remaining staff. Staff shortages also affected prisoners, who would be locked up for longer periods, the ensuing frustration would then be released on staff, aggravating the situation still further. Insufficient staff also affected control and monitoring of bullying and reduced the amount of time which prisoners were able to spend with visiting family members.

Conclusions: Greater consideration should be given to understanding the wider environmental and organisational factors that contribute to poor mental health in prisons. This information can be used to inform prison policy makers and managers, and the primary care trusts who are beginning to work in partnership with prisons to improve the mental health of prisoners.

Text A4

Note on costs

- The annual median cost of incarcerating a prisoner in secure custody in 2004 was about US$28,000 per state prisoner in the United States, US$45,000 in Australia, and US$53,000 in Britain.
- US state prisoners' annual (2001) healthcare costs averaged 12% of total operating expenditure (around US$3,350 per prisoner) while food service accounted for 4.5% and utilities for 3% of total costs.
- Reform of prison health services might reduce costs.

END OF PART A
THIS TEXT BOOKLET WILL BE COLLECTED

Appendix B – Strategy questionnaire

Instructions: For each statement, decide on your response – from "Never" (1) to "Always" (5) – and put a cross (X) in one of the five columns on the right.

No.	Your thinking	Never	Sometimes	Often	Usually	Always
		1	2	3	4	5
1	When I started the test, I planned how to complete it and followed my plan.					
2	I scanned through the reading test before I actually started to complete it.					
3	I translated the reading texts and task into my first language.					
4	I summarised in my head the main information in the texts.					
5	I made short notes or underlined main ideas during the test.					
6	I tried to understand the relationships between ideas in the texts and task.					
7	I used the titles of the texts, tables or figures to help comprehend them.					
8	I used my knowledge of how information is structured in English to comprehend the texts.					
9	I predicted what was going to come next while I was reading the texts.					
10	I analysed what the author meant or tried to say in the texts.					
11	I spent more time on difficult items.					
12	I tried to understand the texts and task regardless of my vocabulary knowledge.					
13	I tried to find topics and main ideas by scanning and skimming each text separately, one after the other.					
14	I tried to find topics and main ideas by scanning and skimming across all the texts at the same time.					
15	I read through the texts and task several times to better understand them.					
16	I related the information from the texts or task to my prior knowledge or experience.					
17	I tried to identify easy and difficult test items.					
18	I was aware of what strategies I was using in the test.					
19	I checked my own performance and progress while completing the test.					
20	I guessed meanings of unknown words.					
21	I took time to review my responses to the items.					
22	I was aware of how much the test remained to be completed.					
23	I knew which information was more or less important.					
24	I was aware of the need to plan a course of action.					
25	I knew when I should read more quickly or carefully.					

1. Profession: ____

2. Specialisation (if any): ____

3. Years working in profession: ____

4. Current workplace *(check ☑ as many as relevant)*

□ Hospital
□ Private practice
□ University/research institute
Other (please specify): ____

5. English language test (*check ☑ as many as relevant)*

□ I have taken the OET before.

- Date (month/year) of most recent test: ____________________________
- Reading sub-test result (A-E): ____________________________

□ I will be taking the OET in the next three months and I allow the OET Centre to provide my Reading sub-test result to the researchers for the purposes of this research.

6. Full name: ____
(required to obtain test results from the OET Centre)

Appendix C – verbal report protocol

Outline of the session

You are going to take an English language reading task. Please try to complete the task to the best of your ability.

Please read and follow all of the task instructions carefully. You have 15 minutes in total to spend on the task itself, and this will be split into 6 x 2.5 minute sections. After the first 2.5 minutes, I will ask you to stop working on the task and to tell me what you remember thinking as you completed that 2.5 minute section of the task. I am interested in what you actually remember about what you were thinking, not what you think you may or should have thought. If possible, it would be best if you can tell me what you remember in the order in which your memories occurred as you worked through the questions in the task. Please talk as much as you can.

When you have finished talking about that section, I will tell you to recommence for another 2.5 minutes. Again, after the second section, I will ask you to stop working on the task and to tell me what you remember thinking as you completed that 2.5 minute section of the task.

We will do six sections in total. I will tell you when we are starting the final section. After the final 2.5 minutes, I will again ask you to stop working on the task, wherever you are up to, and to tell me what you remember thinking as you completed that section of the task. I might ask you some follow up questions, as a way of helping you remember other things about what you were thinking as you completed the task.

I will not talk to you while you are completing each section of the task, and you should not ask me any questions about the task once the timing has started.

As you talk, I will be recording your voice. This recording is for research purposes only.

Do you have any questions about what we'll be doing today?