Summary, Implications and Suggestions for Further Research

*Khalid Addamigh, PhD
King Saud University, Saudi Arabia

Following Widdowson's assertion that, with regard to a study result, that "the expression 'it works' should not mark the end of an investigation but the beginning" (Widdowson, 1983: 36), the objective of this final chapter is to highlight the potential implications of the main results of this study for a variety of L2 domains. It examines the construct validity of MCFGV tests and various L2 theoretical and practical issues, as pointed out in I.3. It will also put forward some proposals for future research.

X.1 Construct validity of MCFGV tests

This section explores the degree of construct validity of the test investigated based on the exploration stated in Chapter IX. Prior to judging the construct validity of a test, one must first specify the theoretical construct or assumption/s behind developing that test, i.e. the variable it is claimed to measure. It has been mentioned in Chapter II that the basic validity question for MCFGV tests can be stated as: to what extent do such tests measure knowledge of vocabulary? However, something that has emerged more clearly than before, as a result of our introspective study of TTS, is that this formulation is actually vague and needs to be sharpened up. Even if a MCFGV test measures lexical knowledge, there are a number of finer distinctions to be made about what precise lexical knowledge is supposed to be, or actually is, measured. Hence, a number of points of view about the validity of an MCFGV test could be distinguished. One might debate whether a MCFGV test is valid if
(a) it measures only knowledge of one meaning of each of the target words. This sort of construct can be seen in Nation’s vocabulary test. In this view, the scores should not reflect anything mentioned under b or c below.

(b) it measures knowledge of some aspect of the target words, which could be their collocation or part of speech etc., not only the words’ meanings. Scores then still reflect knowledge of the target words, but not always knowledge of their meanings.

(c) it measures knowledge of some aspects of vocabulary items used in the test items, but not necessarily the target words, i.e. it could be of the distracter words. The test in this case is not measuring knowledge of the target words only, but it is still measuring knowledge of EFL vocabulary.

Even though one may argue that construct c is too extensive, we think it could still be a valid construct for the test for two reasons. Note that since we were not the test-maker of any of the items, the decision about the construct that the items in the study were supposed to test was in the hands of the teachers and professionals who wrote the items, not us. However, it is possible to gather some indirect evidence on what could, or should, have been the precise idea of vocabulary knowledge that they had in mind (consciously or not).

1- PMT: The items of our PMT were derived from tests that aimed to measure general level of proficiency, i.e. Cambridge CPE and Michigan tests. Makers of such tests aim to include measures of vocabulary knowledge, which proficiency theorists would regard as a component of language proficiency. However, they do not typically say in specific detail what exact form the ‘vocabulary knowledge’ construct has for them. They are usually not interested in checking
any specific words or limiting their scope to lexical meaning only. The underlying aim is simply to check the testees’ level of lexical proficiency (construct c).

2- TMT: The EFL instructors and courses test knowledge of specific words in such tests, chosen to represent the words in the previous syllabus/instruction, so constructs a and b are involved here. However, since the fundamental aim of our subjects’ EFL instructors and courses was to develop the learners’ EFL ability, this aim would still be met if the scores obtained on achievement vocabulary tests were such that those who obtained higher scores had more general vocabulary knowledge than those who obtained lower scores. If this is the case, view c would be appropriate.

Now, we turn from discussion of the precise construct that the test was supposed to test to the evidence of what it does test and, hence, its validity with respect to views a, b or c of the construct. Our study provides three means of validation, of which the third is the major one:

1. Traditional correlational construct validation (as in II.4.5.a). The correlation of subjects’ test scores with TOEFL was highly positive, r=851 and p=.001. However, since any theory of language proficiency would probably regard all three vocabulary knowledge (views a, b and c) constructs, as parts of general language proficiency and so predict high positive correlations of scores on a test of any of the three with language proficiency, this tells us nothing. This sort of approach is too crude to allow us to distinguish between very similar constructs and validate a test for one rather than another. The same argument would apply to the correlation between the scores of our
test and Nation’s. Thus, statistical analysis does not suggest any great damage to validity, perhaps because it reflects a broad picture. Further, as will be explained later, there was no ‘mismatch’ between any significant correlation between TTS and the study vocabulary test scores on the one hand, and any correlation between TTS and Nation’s vocabulary test on the other.

2. Along with our introspective protocol analysis, we necessarily analysed the language of the test items with a view to what clues were available to assist choice (the ‘discourse analysis’ approach to construct validation mentioned in II.4.5). It was found that some of the items clearly would not be valid if we interpreted the construct of our MCFGV test narrowly as view a. The test was apparently designed to examine more lexical knowledge than only words’ meanings. There were items that were not exactly meaning-related, i.e. some items provided synonym alternatives which allowed for the answer to be positively picked on the basis of other vocabulary knowledge, not only meaning. An example is the following item, where collocational probability is the focus:

- David could not convince the bank that he would use the loan for Business _________.
  A. intentions       B. purposes          C. goals             D. aims

We notice that the distractors do not always correspond with the part of speech or other syntactic features of the gap. Hence, they could be eliminated based on that feature. For example, the following item requires an uninflected verb, but one distractor is an inflected verb and one not a verb at all:

- I’m afraid, our insurance policy does not ________ dental care.
  A. scare       B. cover             C. accident            D. insuring
We also found an item that was invalid in all of views a, b and c due to the test construction. The alternatives provided for the following item were, in fact, all correct answers, though the tester regarded only ‘trap’ as correct:

- The hunter forgot where he had put the __________.
  
  A. trap          B. cage         C. bag             D. hook

That was clearly an invalid item, because a testee may select an incorrect answer based on legitimate reasoning, and the introspective protocols supported that. However, it was the only test item, out of 30 items, that had such a problem, which means that test validity was not seriously affected in this way.

3. Our main validation comes from the think-aloud access direct into the test-taking process. The TTS evidence allows us to assess what test-taking behaviour would be consistent with, or damage, validity with respect this or that construct. The above different definitions of 'what was the underlying objective or variable intended to be measured by MCFGV tests' (the construct) would lead to different views of what strategies lead to invalidity. It was found that the most effective strategy that guided generally to the right responses when the target word was unknown was strategy e36, elimination using pragmatic likelihood between the meanings of distractors words and the meaning of the stem, see IX.4 for details. This strategy does not damage the construct validity of the test with respect to the construct defined by view c, as it still distinguishes between EFL test-takers according to their amount of vocabulary knowledge. However, it makes the test invalid in sense a or b. Moreover, if the strategies used to help in arriving at the correct answers are strategies that use other lexical knowledge, such as e38, eliminating an alternative using part of speech, or e44,
selecting an alternative using L2 wording of a collocation, and are effectively used by
testees with higher Nation test scores, then validity in sense c is also not damaged. The
test still differentiates between test-takers with greater lexical knowledge and those with
less, since the test responses, in this case, involve both knowledge of the tested
vocabulary and ability to use lexical knowledge-dependent TTS to tackle the test items.
In this respect, TTS make the test valid according to view c of the construct, but not
necessarily according to view a or b.

However, the analysis revealed that there were some cases in which the test-takers
selected correct answers, even if they did not possess any lexical knowledge, based on
strategies that are not lexical-related, so they lead to invalidity in all the proposed senses
of the construct (views a, b and c). These strategies can be grouped into two categories:

1. Non-linguistic strategies, e.g. e49, selection of a letter that had been less
   frequently chosen in the other items.
2. Linguistic strategies that were quite independent of lexical knowledge, e.g.
   e37, elimination using general grammatical rules.

Clearly the last strategies mentioned would (under any definition of the construct) cause
harm to the construct validity of any vocabulary multiple-choice test. However, the
overall validity of the current MCFGV test investigated was not particularly badly
damaged for three reasons:

1. Generally, any multiple-choice test format, for any language skill or sub-skill,
   would be open to the exploitation of non-linguistic strategies. It is difficult for
   a test-maker to avoid the use of such strategies: i.e. this is a universal validity
   problem with multiple-choice tests.
2. The instances of cases getting unknown items correct by using such strategies formed a rather limited part of the test-takers’ overall test-taking processes in our test.

3. The quantitative analysis does not show a positive significant correlation between those strategies and the test scores.

A final more general argument about validity based on the TTS is as follows, using quantitative analysis. If a TTS correlates positively with scores on the study test, it also correlates positively with Nation’s test and the same is true of the negative correlations. It would damage validity if there was a TTS correlating negatively with Nation’s scores and positively with the study test scores, but there is not. The only TTS that has a relationship with the study test scores, but does not even get close to significance with Nation or TOEFL is e47, selecting an alternative using a lexically dependent grammatical rule. That is probably because it involves grammatical knowledge as much as lexical, though it falls within view b of the construct.

In conclusion, the qualitative analysis has led us to a clearer understanding of the different possible interpretations of the construct of vocabulary knowledge in an MCFGV test. Further it showed that there were some cases where the answers to MCFGV test items were unknown and the testees could obtain marks based on exploitation of strategies using non-lexical knowledge. However, the test construct validity, if interpreted in sense c, was not badly damaged and the test still robustly distinguishes between test-takers according to their lexical knowledge. The discussion in this section also supported the claim discussed in II.4, that test-takers could select an incorrect answer by using legitimate reasoning, or select a correct answer without having any actual related lexical knowledge, by applying certain strategies.
X.2 Theoretical implications

The conclusions drawn from the present study have relevance for a number of L₂ research theoretical concerns. These which will be examined in this section with regard to three main aspects: the implication for strategy research, Krashen’s theory and the PPP model.

X.2.1 The TTS types and frequencies

One of the major contributions of the present study to L₂ research is simply the detailed qualitative exploration of the test-takers' abilities to cope with problems in multiple-choice vocabulary test items, arising from the verbal protocol analysis. That qualitative analysis, combined with evidence in Chapter VII, reveals that 62 different ‘types’ of test-taking strategies were used, compared with the 23 in Nevo’s (1989) list and the 17 relevant strategies in Anderson’s (1991) list, both of which were on reading comprehension TTS. The TTS revealed in this study were used, with different frequencies, to handle the MCFGV test as a whole or to solve particular L₂ problems in the test items. The remarkable number of TTS types used by subjects reflects all the strategies found in the data gathered, i.e. not every subject necessarily used all those strategies. Quantitatively, we find that the average repertoire of different types of TTS used per testee was 29 (SD = 3.65), which is slightly under half the maximum possible size of TTS repertoire found in the entire data. The largest TTS repertoire was used by a ‘low’ proficiency subject, being 38 different types, i.e. 61% of the total types coded. By contrast, the test-taker who used the smallest TTS repertoire was a ‘high’ proficiency test-taker, using only 19 different types, i.e. 30.65% of the 62 types in the TTS inventory. We also find that all the EFL test-takers possessed a range of abilities to use TTS with an average of about 98 TTS tokens per testee (SD = 14.01) giving an average of 6.3 per MCFGV test item per testee. Due to the nature of the study task, most of the TTS used were related to L₂ lexical knowledge. However, some of the TTS
choices made went beyond strategies at the level of lexical knowledge. There are cases where the test-takers decided on their lexical responses by considering the test context, e.g. strategy e49, selecting a letter less frequently chosen in the other items, or a wider L2 aspect of the context, e.g. e37, elimination using general grammatical rules, or L1 context, e.g. e43, selecting using the L1 wording of a collocation. Since the domain that was investigated by the study, MCFGV test items, has not been, to our knowledge, explored by any previous researcher, some of the TTS discovered in this study are new to the field. For example, the strategies revealed in category c, on how the testees handled the gap in the multiple-choice test items, are thought to be an original contribution. We find that eight strategies were used to handle the gap in the stem. The most common strategy was simply to skip the gap while reading, c17, which was used in 35% of the total items tested. Strategy c20, using an L1 or L2 filler word, came next in frequency, being used in 31% of the total items tested.

X.2.2 The TTS Taxonomy

When classifying the TTS found in the data into subcategories, we noticed that previous literature on TTS listed the strategies in a way that is not systematic. For example, in Nevo’s reading comprehension TTS checklist, reviewed in II9.1.1, strategies 1 and 3 are related to choosing the answer based on the passage, whereas the strategy placed between them is “blind guessing”. In Alderson’s study, reviewed in II.9.2.1, strategy 11 is about “changing an answer after having marked one” and strategy 13 is “stopping reading the options when the testee reaches the answer”. Logically, however, one chooses an answer first and then might attempt to change it. In this study, the different types of strategies are grouped into a seemingly logical system according to the particular phase of tackling a test item that they served. The categories are, for the first time to our knowledge, ordered according to the 4 hypothesised steps of answering a MCFGV test (as explained in V.4.2). This is substantially confirmed by the data:
• Strategies for managing the test as a whole, which may occur before, during or after the strategies associated with individual items, categories ‘b’ onwards.

• Strategies for stem comprehension.

• Strategies for handling the gap.

• Strategies for handling the alternatives.

• Strategies for choosing the answer (positive and negative line of attacks).

• Strategies used after choosing the answer.

This system of classifying the TTS could be a useful model to adopt in a similar study, for exploring strategies in a multiple-choice test.

X.2.3 L₂ proficiency and TTS

The study revealed that ‘high’ proficiency test-takers, of both general and lexical language proficiency, made less use of strategies than ‘low’ proficiency test-takers, in both the number of TTS tokens and the repertoire of types of TTS. Category e particularly showed a significant difference between the two groups. This is probably because the low proficiency subjects needed to use the strategies in this category more than the high proficiency testees, since they were likely to face more instances of inability to access the target words from their MLs, because of the small amount of vocabulary they had. However, the results of other studies that examined the impact of the subjects’ level of language proficiency on the exploitation of process strategies are inconsistent and some of them revealed that higher proficiency subjects are more strategic. Olshavsky (1977) found that L₂ ‘good’ readers used reading strategies more frequently than ‘poor’ readers in the EFL reading comprehension process. Anderson (1991) found that high and low proficiency students used the same strategies in textbook reading comprehension as well as in a reading test. The findings of Singasiri’s (2001) study in this respect are consistent with those of the present study. She found
that low proficiency subjects used more strategy tokens and types than high proficiency subjects in EFL writing and reading. The reason for the greater use of TTS by low proficiency subjects in this study is probably because the high proficiency test-takers faced fewer problems in the test, since they were more likely to be able to access the required lexical items from their ML and select the right one automatically. This would mean that the low proficiency test-takers had more problems in the test simply because the number of instances where they did not know the correct response was greater, so they needed to use some more conscious strategies to solve the lexical problem. The following section shows more details of the difference between the two groups of subjects.

II.2.3.1 L2 proficiency and individual TTS

The difference between the high and low general/lexical proficiency testees in the number of TTS tokens, was particularly significant with some TTS. It was found that the low proficiency testees used significantly, or near significant, translation and rereading/repetition, i.e. b15 (with both types of proficiency), b16 (with the general proficiency), d34 (with the lexical proficiency) and d35 (with the general proficiency), to handle the stem and the alternatives significantly more than high proficiency test-takers. Rereading and repetition also characterise lower proficiency readers generally (Olshavsky, 1976/7; Upton, 1997). The low proficiency test-takers also, not surprisingly, stated problems in the alternatives more than the high proficiency group, d32. The high proficiency testees showed more evidence of inferencing a meaning for the gap, in L1, which is again consistent with high proficiency student reading process behaviour (Romero, 1999). In contrast, the low proficiency testees were significantly more likely to skip the gap, c17, than the higher proficiency subjects. Indeed, inferencing the meaning of the gap word, as was done by the high proficiency testees,
needs stem comprehension, which low proficiency subjects are more likely to lack than are high proficiency subjects. In category e, high proficiency subjects used e44, selecting L₂ wording collocation, significantly more than low proficiency testees, who tended to use L₁ wording collocation, e43, significantly more, which probably connect to their use of translation above. The low proficiency subjects also relied significantly more on, e45, selecting an alternative using pragmatic meaning likelihood in relation to a word in the stem. The low lexical proficiency subjects tended to use e38, elimination using part of speech, more than the high proficiency subjects. This is possibly because knowing a word’s part of speech represents a lower level of word knowledge than knowing its meaning, so the lower lexical proficiency subjects tended to use it to compensate the lack of knowing the meaning of the target words. They also tended to use e48, considering the letter of the preceding item’s answer, which can be used without relying on any knowledge of the alternative words provided. The low general proficiency test-takers were likely to use e56, random choice, more than the high proficiency subjects. Again, this is perhaps because the high proficiency test-takers faced fewer problems in the test, since they were more likely to retrieve the required lexical items in their MLs automatically (route a in Figure 2.4). The other strategies did not show significant difference between the two proficiency groups probably because most of them were not used much. If a strategy is little used, it has less chance of showing a difference between groups. On the whole, there was a similar picture of what differentiations a high from a low proficiency testee in both the general and lexical proficiency. Apart from the strategies, b16, d34, d35, e38, e48, e54 e56 and f58, the significant and near significant differences found between the TOEFL groups were the same as those found between the Nation’s test groups. The explanation for this similarity perhaps because most of the subjects who were highly proficient on TOEFL were also highly proficient on Nation’s test. See VIII.3 for further details.
X.2.4 Differences in TTS between the PMT and the TMT

The analysis reveals that the average number of strategy tokens used in the PMT items was 6.28 tokens per test item, SD = 6.9, whereas in the TMT items it was 6.24 per item, SD = 8.5. This is a minor non-significant difference. In relation to the total number of tokens within different TTS categories, the major differences (more than 5%) appear in two categories, b, the stem comprehension strategies and e, selecting a response.

In category b, the number of tokens in the PMT items is 5.4% greater than the number of tokens used in the TMT items. This is because the use of b12, stating problematic vocabulary, occurs in the PMT significantly more than in the TMT, 7.34% and 3.04% respectively (a difference of about 140%). This might indicate that the international PMT stems contained more unfamiliar vocabulary than the local TMT stems.

In category e, the TTS tokens in the TMT items outnumber those in the PMT by 6.6%, which was the greatest difference in number of tokens for any of the TTS categories. Indeed, this difference alone is significant, t = 2.182 and p = .037. Strategy e36, elimination using pragmatic meaning unlikelihood, is one of the strategies that shows a particularly strong significant difference between the PMT and the TMT, t = -.3.44 and p = .002, being used about 50% more in the TMT items. This indicates that the alternatives provided by international professional test designers were more meaning related and more distracting than those made locally. Another strategy that is used significantly more with the TMT than with the PMT is e45, selecting an alternative using pragmatic meaning likelihood in relationship to a word in the stem, 15.45% and 9.27% respectively, t = -2.814 and p = .008. This is possibly because the PMT items often involved synonym alternatives whereas the alternatives provided for the TMT items were, more often, totally different in meaning. This suggests that the test-taker could usefully use strategies, such as e45.
By contrast, some strategies were used significantly more with the PMT items. One is e44, selecting an answer using L2 wording of a collocation, which was used in 6.82% of the PMT items and in 4.96% of the TMT items ($t = 2.329$, $p = .027$). The greater use of this strategy in the PMT is, possibly, because the number of test items that examined knowledge of lexical wording of a collocation was greater in the PMT than the TMT, see for example test items number 1, 4 and 5 in form A, UQU version, Appendix D.2. Another notable strategy is e56, selecting an answer based on random choice, which was used in the PMT with 7.54% of the test items and in the TMT with 6.18 of the test items ($t = 2.171$, $p = .038$). This might indicate that the PMT items had a tighter design, limiting the possibility of use of other TTS. This, in turn, increased the use of irrational random guessing, an expected behaviour in a more difficult multiple-choice test. That might explain why, although in the current study the two sets of items were picked in a parallel way (see IV.5.10 for details), the average of scores earned from the TMT items was greater (70.5%) than those from the PMT (66.7%).

With regard to the differences in the repertoire of different TTS types used to tackle each of the PMT and the TMT items, the analysis reveals a very minor non-significant difference in the total number of different strategy types used, 556 for PMT and 548 for TMT, giving averages of 17.4, SD = 2.56, and 17.1, SD = 2.23, per testee. The analysis also shows that the differences between the categories in the total number of different TTS types used is minor and non-significant, less than 5%.

Overall, apart from the effect on the way in which unknown answers are selected (which is likely to be related more with the choice of the alternatives provided) the test-type had only minor effects related to the precise way the items were constructed. The construction of the two sets of items, of course, was not controlled in this study. Hence,
the two types of test stimuli varied in terms of how the alternatives were chosen, e.g. synonyms or not, different in lexical grammar or not, etc.

X.2.5 Subconsciously acquired L₂ vocabulary

It is worth noting that there were some cases of successful attempts in which the test-takers selected the correct lexical items because they thought they 'sounded nice' or because of a ‘feeling’ that they were the correct words. They did not actually indicate that the meaning of the alternative chosen was known; for example, see e.52 in VII.1 or case number 11 in IX.4.4. This level of 'feeling' might be consistent with Krashen’s (1988) theory that the subjects had acquired those lexical items subconsciously. His 5-Hypothesis Acquisition-Learning Theory suggested that adult SL learners have two ways of developing an L₂ competence: (a) subconscious acquisition, which is identical to the children’s process of acquiring their L₁, and (b) learning, which is a conscious process that results in monitoring L₂ rules and systematic knowledge. Thus, one of the features that distinguishes between ‘learning’ and ‘acquisition’ in that hypothesis is the conscious level of awareness in obtaining language knowledge. Our qualitative analysis also shows that in some instances the first instructive response was replaced by a second incorrect monitored choice. This corresponds with Krashen’s theory that acquired knowledge is superior to learnt, which further agrees with the advice in test-taking guides. This result is not in support of testees changing their first choice and supports Alford; (1979) view that second-thoughts are likely to result in wrong answers.

X.2.6 The PPP model

This study does not aim to examine how robust the PPP model is (outlined in Chapter II.6). However, to our knowledge, the implications of this model for the test-taking process have not been tested before. Our findings, Therefore, may throw some light on this framework. Although it is bay and the scope of this thesis to conduct a full-scale
analysis, there is some evidence to support both the direct route from presage to product, route ‘a’ in our model (II.6.3), and the indirect route from presage to product through process, route ‘b’ in the model.

First, with respect to ‘a’, the qualitative analysis (Chapter IX) summarised in Table 9.2 shows that 15 of the item cases analysed were answered by direct access to the response without the use of strategies (in our conscious problem solving sense of the term, Chapter II.5). Arguably, these represent instances of route ‘a’. We further find that all these items were answered correctly and that 10 of them were responses from done by the high proficiency group. This, therefore, provides some limited descriptive support for the hypothesis that higher proficiency test-takers, more often than those with a low proficiency level, use direct access to the answer from their ML, and by doing so, get the answer correct. This indicates that a presage variable (high and low language proficiency) is directly connected to a product one (getting the items answered right).

Furthermore, with respect to ‘a’, if we use the Partial Correlation of TOEFL scores with the study test scores, ‘partialling’ out (i.e. eliminating the effect of) the scores obtained in combination with 15 TTS that showed significant relationships with the study test scores, a significant correlation of $r = .638$, $p = .006$ is still obtained. This shows that there is a strong relationship between the subjects’ general proficiency and their test performances independent of the effective use of strategy. A similar result is found for the Partial Correlation of Nation’s test scores and study test scores ($r = .694$, $p = .002$). Consequently, once again, the existence of a direct presage-product (route ‘a’) is supported.

Second, with regard to route ‘b’, the analysis shows a significant negative correlation of the total number of TTS tokens both with study test scores ($r = -.471$, $p = .006$) and with
the testees’ general and lexical EFL proficiency level (general proficiency $r = -0.591$, $p < 0.01$, lexical proficiency $r = -0.520$, $p < 0.01$). In other words, more proficient test-takers are likely to use fewer TTS tokens and a smaller repertoire of different TTS types and they also obtain high test scores (more details are in VIII.2). The qualitative analysis, Chapter IX, shows a range of precise ways in which language proficiency level has an influence on test scores via TTS, i.e. route ‘b’. For example, with a high language proficiency level, a learner will be a testee who does not rely on familiar strategies that are used by poor readers, such as rereading or translation, to understand the stem in stage 1; can inference meaning for the gap and not skip it, in stage 2; does not encounter many problems with the alternatives or need to repeat or translate them, stage 3; and in the crucial choice stage uses L2 collocational wording rather than L1 collocation and does not use strategies with no sound linguistic basis, such as relying on eliminating unknown alternatives or selecting odd alternatives. Accordingly, it can be said that the Presage category, embodied in the L2 proficiency level, would also indirectly influence the test Product, i.e. the test scores earned, through the Process categories, i.e. the TTS used. More notably, in our study, 11 out of 15 specific TTS that correlate significantly with study test scores also correlate significantly with general and lexical EFL proficiency. This all means that the Presage factor (L2 proficiency) influences the Process factors (TTS use) and so the test scores product, which corresponds with rout ‘b’ in Figure 2.4.

In conclusion, though, of course, cause and effect cannot be demonstrated conclusively by a correlation study, the result is consistent with the argument that the three categories of factors of the PPP model function together in a successive interrelationship and each has a sequential effect on the other, with some being more direct in their effect (as in the relationship between language proficiency and product or between language proficiency and TTS) and others being indirect (as in the relationship between language proficiency and product or between language proficiency and TTS).
proficiency and product through TTS). Applying this result to the sequences hypothesised by the PPP model, it can be said that our findings support the conceptual framework of that model, as the presage variables influence both the product and the use of certain process strategies, which have an effect on the product. This, in turn, means that the presage factors also have an indirect effect on the product through the process factors.

X.3 Practical implications

The practical implications of the findings of this study can be demonstrated in a number of ways. This section will focus on TTS training, test construction and strategies research methodology.

X.3.1 Test construction

The practical side of research on test validity is to improve test construction with the aim to avoid providing the correct answers to the testees who use TTS through flaws in the test construction. A better understanding of test item problem management by testees taking L2 tests is a good basis for test designers to improve their assessment measures. This is in line with Cohen’s (1998) view of the consequences of such information. He states that

> The insight gained from looking at the test taking strategies used by L2 learners can help both to improve the assessment instruments themselves and improve the success the learners have in responding to these instruments. (Cohen, 1998:215).

The main weakness found in the test items involved in our study test was that some of the distractors provided were inactive, especially in the TMT items. This allowed the test-takers to use the elimination strategy, e36, in more than one quarter of the total items tested. The qualitative analysis, in Chapter IX, shows that this was an effective strategy. To overcome this weakness, the test constructors could make the distractors more active by looking at the common features of the alternatives most attractive to the
test-takers. In this respect, it was found that when testees did not know the answer for a MCFGV test item, they widely used the $L_1$ and the $L_2$ wording of a collocation to select their responses, e43 and e44. The most obvious instance of an $L_1$ collocation distractor that attracted most of our testees who were not sure about the correct completion was ‘exchange’ in the following Cambridge test item:

- If you’re bored with this TV programme, why don’t you ________ channels?
  
  A. hop       B. press             C. exchange            D. switch

In Arabic, the collocation is $بَيْعِّي ة$ القناة, where the word $بَيْعِّي ة$, means both ‘change’ and ‘exchange’, so many of the subjects went for ‘C’. Option ‘A’, hop, was not active for Arabic speaking testees, because it does not collocate at all with the meaning of the stem in the $L_2$ or in the $L_1$. It could, however, be active for speakers of other languages, Japanese test-takers for example. The testees also tended to select their answers based on pragmatic meaning likelihood in relation to a word in the stem or to the stem as a whole or part of it, e45 and 46. Hence, distractors with lexical features that allow this would be more active, and would improve the quality of the test construction. The testees also used the whole test context to decide on an unknown correct answer by choosing a letter that had been less frequently chosen in the other test items, e49, or that chosen for the preceding item, e48, which do not involve any lexical knowledge. Test-makers should be aware of these strategies when placing the correct completion among the distractors. As can be seen from the above examples, although the main scope of this investigation was the test-taking strategies used in multiple-choice vocabulary tests, and although its results can be generalised only to the population studied, our findings can provide test constructors with useful information on how to construct better multiple-choice test items.
X.3.2 TTS training

The results of this study cannot be taken as a basis for prescription about what TTS to teach and how. However, any research on test-takers’ TTS may provide constructive information on the successful individuals’ TTS and so help in the construction of courses and compilation of materials used to train testees. This is because

Students provide the first input into instruction in the form of learner strategies and teaching consists of adapting to this input. (Hosenfeld, 1977: 52).

In any case, there are arguments both for and against the appropriateness of providing test-takers with knowledge on how to take tests. The basis for this disagreement stems, probably, from the way that TTS training is interpreted. Opponents of TTS training suffer perhaps from a narrow understanding of such a project, limiting to passing on information about the actual knowledge being tested. The contrasting argument relies perhaps on a broader interpretation of teaching TTS as involving different objectives, such as raising learner awareness of the importance of TTS and encouraging learners to use TTS to support their linguistic abilities. A view that is in favour of such training courses, claims that

Many students are poorly prepared for test-taking strategies and that the lack of "test-wiseness" is often responsible for test scores which do not accurately reflect the students' knowledge. (Scruggs and Mastropieri, 1995: 1).

On the other hand, others oppose such courses and materials. For instance

The UCLES/British Council ELTS adopted for some years the policy of not releasing specimen papers, merely descriptions of the test tasks on the argument that it was not particularly important for candidates to know how to pass the test, rather than they be able to perform successfully in the target-language use situation. (McDonough, 1995: 105)

Some of those who do not favor TTS training might argue that such training would enhance the invalidity of the test by training testees to exploit to the full any way of arriving at the correct answer other than by knowing the information the testers want to test. However, in reply to this, the current study establishes that it is a matter of fact that
some testees have already developed some successful TTS to handle the test effectively or to compensate for deficiencies in their L₂ knowledge. This would mean that all test-takers should have the right to such strategies, in order to be equal in the ability to perform effectively in the test, as well as have an equal chance to earn marks on difficult test items. Any invalidity of the test arising from the use of TTS would be the same for all, at least. The remedy is for the test constructors to make sure that the tests are made so that the only TTS that will be effective are ones that indirectly rely on the knowledge that the test is meant to measure (see X1.7 and X 2.1).

Our study shows a lack of TTS training with Saudi Arabian students in their schools, universities and other courses. Only one testee indicated that he had been trained in using TTS, but when he was asked about that it appeared that all he had received was some information on the TOEFL sections and their test formats. It would, therefore, be useful for Saudi students to be trained on how to use TTS, to help them make the most of their knowledge. We think that the introduction by the Ministry of Education, of two academic subjects at the end of middle school stage (age 15) and at the end of secondary school stage (age 18), would be suitable for such training.

When deciding what test-taking strategies to give training, it should be considered that not all the strategies should be given the same treatment. Some TTS are clearly not desirable to teach. An example, taken from the data provided by the current investigation, is that of skipping the test instruction or changing the first answer for an uncertain response. There are, perhaps, not desirable TTS to teach. Other strategies, however, like reading the whole stem before examining the alternatives, elimination using pragmatic meaning unlikeness, using a lexically dependent grammatical rule or the L₂ wording of a collocation are not only useful in a vocabulary test but are also amenable to training. This results found in our study, therefore, contribute useful input for courses that aim to train test-takers in how to take any test battery that involves
MCFGV test items. For example, since our study uses samples of the vocabulary sections from the Cambridge and Michigan Universities test batteries, the output of this study might be helpful in contributing useful input to the courses being run on how to take these test batteries. The findings also could be useful to the material being produced on how to deal with the two batteries mentioned above (e.g. Sharpe, 1982; Witt, 1997) as well as that which examines TTS in general (e.g. Orr 1994; Boone 1996; Jasmine 1999). Our findings could be used to draw attention to the effective strategies and provide advice to students on how to avoid the ineffective ones. Such materials could emphasise the positive use of the $L_2$ wording of a collocation and the negative result of using the $L_1$ wording of a collocation in an $L_2$ test context, since the two languages differ markedly in collocation.

**X.3.3 Methodology and other strategies research**

Although the main focus of this study was not the methodology for investigating process strategies itself, it is useful to mention any constructive points we came across in this respect. It is felt that the methods of data collection and analysis utilised in this study were successful in accomplishing the objectives of the study. These methods, therefore, can be confirmed as effective for exploring human language processing. To add more validity to the findings of this study, the data collection was based on triangulation of three complementary instruments. See IV.5.1 for further details. This multiple exploratory data collection method seems to reveal rich information about the targeted mental processing. Analytically, this study demonstrated the value of investigating the multiple methods simultaneously, since the information was treated at the same time to provide a clearer picture. The feedback from those three instruments, therefore, was neither coded nor analysed independently. Such an approach is expected to reduce the constant threat to internal validity. The above method and analysis can be recommended as a model for similar process-based research.
The preliminary review of the types of EFL lexical test format, Chapter III, faced the problem that the vocabulary test items obtained for the preliminary study could not readily be classified according to any traditional broad-spectrum taxonomy (short answer, multiple-choice, true/false questions etc). See III.4 for more details. Thus, a new scheme was developed to categorise these vocabulary test items in a more discriminative way. This new test item taxonomy was based on a three dimensional perspective: (a) the test task itself (definition, gap filling, etc.), (b) the sort of support given for the answer (yes/no, multiple-choice, etc.) and (c) the context in which the target word is presented (isolated, sentence, etc.). This new classification scheme can, in fact, be considered as a useful model for future work of this type, to help develop a neater discriminative taxonomy for classifying test items.

X.4 Suggestions for further research

To help TTS research emerge from its current narrow boundaries, this section suggests further issues that need to be looked at in the test-taking field as well as in other related areas. Some of the issues that relate to the data collected and methodological concerns will be dealt with first.

(1) Although none of the aims set for this study was to investigate the influence of the L₁ on L₂ processing, the marked presence of the L₁ in the processing of the test cannot be overlooked. In a protocol analysis of L₂ writing, Woodall (2002) finds that two factors influence the duration of L₁ use in the L₂ writing process: the duration increases with (a) more difficult tasks and (b) less language proficient subjects. The data collected for the present study shows that the subjects, including the high proficiency subjects, made extensive use of the L₁, Arabic, in processing the L₂, English. This can been seen clearly from the marked use of translating the stem into the L₁, b16, even though the
meaning translated was simple for their level of English and could easily be handled in the L2. An example is the following test item:

- Many people were ________ about the way the criminal escaped from jail.
  
  A. curious        B. guilty              C. innocent            D. correct

In an introspective protocol, a subject read the stem in this way: Many people were ..

many people were .. about the way the criminal escaped from jail .. how the criminal escaped from jail.

Another example is:

- The girl's bedroom was divided in half by a ________.
  
  A. behalf         B. partition              C. strait            D. spine

A subject handled the stem in this way: The girl's bedroom was divided in half by space .. ~ the girl's bedroom was divided in half by ..

In line with this finding, many studies use think-aloud and find a strong presence of the L1 in L2 processing. Upton (1997), for example, investigated the roles the L1 and the L2 play in the L2 reading strategies of ESL readers. His study revealed that L2 subjects had a greater tendency to try and work out text and sentence meaning by translation concepts they understood into L1 (Upton, 1997:18).

Some experts have re-evaluated positively the presence of the L1 in the L2 classroom (Cook, 2000). However, the question that might be posed from our findings is, perhaps: ‘were the subjects encouraged to switch to the L1 when processing L2 text more than usual because they were doing the verbal protocol?’ In other words, did the data gathering method affect the L2 context processing, as the subjects tended to voice their
thoughts in the L₁, although the thoughts themselves might mentally have taken place in the L₂?

(2) When considering the L₂ language proficiency measure in this study, it was noticed that there is an ambiguity in the previous literature in the way studies identify their subjects' levels of proficiency. The problem lies in what proficiency measure each study uses to label the subjects as being in more/high and less/low proficiency groups. Some studies based their decisions on general standardised proficiency tests (Purpura, 1997), others on measures for a particular language skill, reading for example, (Anderson, 1991) and some may have used teacher-made achievement tests (Ahmad, 1987). In addition, some studies may leave out middle proficiency students so as to clearly distinguish the two distinct groups (as in this study), whilst others do not (Bin-Ghali, 2001). This results in some doubt over whether the 'high' and 'low' levels of proficiency in one particular study correspond with the 'high' or 'low' levels of proficiency in another study. In other words, what is labelled as a 'high' proficiency group in one study could be similar, in the absolute level of language proficiency, to what is identified as the 'low' proficiency group in another study. This causes difficulty in comparing findings on the impact of proficiency level on strategy use. There is, in my opinion, a vital need for agreement among researchers on the use of an international proficiency test. I would suggest the University of Cambridge test battery for this purpose, because these tests are widely recognised and are relatively readily obtainable. This would provide more standardised information about the subjects' levels of language proficiency, allowing for valid comparisons between studies in this fundamental factor.

(3) Another issue is that few, if any, conclusions can be generalised from the available TTS literature on strategy use in L₂ tests because of the limited number of studies in this
area, as already mentioned in Chapter I. Nor is it be possible to generalise the findings of this study to test-takers who differ in level of schooling or language background, since this study was conducted in the context of university level in Saudi Arabia. We are of the opinion, therefore, that a replication of this investigation on other subjects from different backgrounds would uncover further information and help to identify whether a different sample of test-takers would produce the same outcomes as those obtained from the test-takers in this research. There is another inevitable limitation in terms of the area investigated. This study was concerned only with TTS used in vocabulary test items. Other L2 skills, such as writing and speaking, and sub-skills, such as syntax, remain to be explored in order to widen TTS based insights into how well tests work in achieving their objectives.

(4) This study has also a limitation in terms of the variables that might influence what we are focusing on. This arises from the fact that the choice of the TTS used and success in using them can be influenced by a variety of factors. The current study concentrated on three independent variables: the test-takers' general language levels of proficiency, their lexical levels of proficiency and two clusters of different test constructors. Further research is needed to explore the effect of other factors that might influence the test-takers' choice and use of strategies, for example, differences in strategy use between speed and power tests or between different test formats. It would also be interesting to investigate the connections between the test-takers’ performances in different areas of language skills. For example, are 'strategic' test-takers in multiple-choice vocabulary tests items also 'strategic' test-takers in multiple-choice reading comprehension tests?

(5) Although the main focus of this investigation was the test-taking strategies used in multiple-choice vocabulary tests and its results can be generalised only to the
populations studied, it provides baseline information for further exploration of the
relationships between test taking strategy use and L₂ test performance success. A project
prompted by the findings of this study would be to investigate in depth the effects of
training in the use of the more successful TTS. Among the TTS used in the vocabulary
test items, elimination using pragmatic unlikelihood strategy seemed, in general, an
effective method in choosing out the correct unknown answer. An interesting focus of
future research could be to investigate the consequence of teaching this test-taking
strategy, using experimental and control groups or one group with pre and post-tests.

(6) Finally, projects to train test-takers should be preceded by exploration of the types of
test-item format used by test-makers. This study investigated the commonest test format
types for testing vocabulary knowledge and found that multiple-choice fill-in-the-gap is
the most common type used by Saudi EFL test-makers; details are in Chapter III. Other
researchers may similarly investigate the types of test formats for other language skills
and in other countries to come up with a wider picture of the types of language test
formats used globally.