The Construct Validation of Dicto-Phrase

Hameed Zahedi
Kerman University
&
Sanaz Laleh-Parvar
Azad University of Shiraz

Abstract
Dicto-Phrase, as a test of listening comprehension, requires test takers to reproduce the gist of selected propositions of language. The purpose of the study was to discover what kinds of linguistic knowledge as well as strategies and cognitive processes candidates make use of while performing on Dicto-Phrase. Five upper-intermediate participants were asked to think aloud while performing on the test tasks. Right after that, they answered related retrospective questions. Based on the transcribed protocols, task analysis, and the objective of the study, a coding scheme was developed for extracting common patterns of performance.

The findings indicate that the participants used intelligent guessing, that is a compensatory strategy, frequently. In doing so, they drew upon topical knowledge and appealed mostly to immediate text rather than pretext or post-text. Whatsoever, they looked for more linguistic clues than discourse ones. Random guessing was also used but much less frequently with no significant gains for the test takers. One more strategy used very frequently was rereading strategy although in most cases revision was not considered.

Dicto-Phrase was found to diminish the role of STM in making responses, boosting the role of LTM due to allowing for responses made on the basis of different levels of knowledge. In fact, the procedure was shown to be capable of recognizing sufficient (comprehensible and acceptable) responses as well as exact ones as correct. Otherwise two thirds of the sufficient responses would be lost. Consequently, the
validity of the test scores representing the listening comprehension construct could be expected to be higher than other measures of listening comprehension recognizing only the exact answer.

**Key Words:** Dicto-Phrase, Dictation, Paraphrase, Response validity, Construct validity.

**Introduction**
Dictation-Paraphrase or Dicto-Phrase (D-P) is a test method developed by Zahedi (1997) to assess global listening comprehension in which test takers are expected to concentrate on the content of the idea units (See the Appendix for a sample D-P). In other words, D-P is designed so that examinees are given the chance to reproduce the propositions of language in their own words. D-P requires that test takers complete blank spaces that represent selected idea units in their answer sheets with the gist of the meaning of the missing proposition (vs., the exact replacement of the lexical phrase of the original text). Each blank space corresponds to one single unit of meaning, i.e., one idea unit. There is a question word in front of each blank space to refresh test takers’ minds. Although D-P seems to allow for different levels of knowledge (Zahedi, 2001), investigation is needed to discover what it really measures which is a question of construct validation.

**Dicto-Phrase, vs., other formats**
*D-P, vs., oral cloze.* In oral cloze, irrespective of which version, examinees are aurally exposed to the material up to each blank space and are then given the opportunity to guess the missing word (Oller, 1979). Accordingly, the differences between the two procedures are as follows: (a) In oral cloze, test takers are required to guess the missing word which seems a short-term-memory (STM) challenge. In contrast, D-P involves examinees in a different process like ‘gist recall’ (vs., exact recall) of the missing materials after they have listened to the whole text. This appears to challenge the long-term memory (LTM). (b) In oral cloze, the blank space may only include one word as is the case
with all the versions of the cloze procedure. However, in D-P, a single word, a phrase, or a sentence may be deleted; what counts in D-P is the idea unit, not the lexical unit. (c) Oral cloze techniques may be used with non-literate populations (Oller, 1979), whereas D-P is not developed for that kind of population.

D-P, vs., partial dictation. In partial dictation, “the portions of the text that are missing in the printed version are the criterion parts where the examinees must [simultaneously and exactly] write what is heard” (Oller, 1979, p. 285). Accordingly, test takers are required to replace the blank spaces with exact words, whereas D-P allows for acceptable or comprehensible reproduction of the missing portions. Besides, in partial dictation, there is a pause right after each missing material is read out so as to give test takers time to fill in the blank. However, in D-P, examinees are allowed to see the printed form after they have listened to the whole materials. Similar to the argument above, examinees performing on partial dictation and D-P seem to have to rely on STM and LTM, respectively.

Previous research on Dicto-Phrase (review)
Zahedi (2004, 2005) found that performance on D-P was easier than that of the listening section of the TOEFL (which uses the multiple choice format) and more difficult than that of the partial dictation. As to the concurrent validity of the new procedure, D-P revealed between 40% and 50% overlap variance with the listening section of the TOEFL which was the outside criterion for D-P. Group validity was also established for D-P through showing discrimination power among different-proficiency groups.

A review of methods and task types tapping LC
In order to explain the motivation for the development of the new test method, some of the current test formats used to assess listening comprehension (LC) is reviewed below.

Multiple-choice test (MC). Multiple-choice test format is undoubtedly one of the most widely used item types. The MC is favored
by both testing organizations and classroom teachers because it provides rapidity, and economical scoring (Hughes, 1989), ease of administration and scoring (Sax, 1989), as well as broad content sampling, high score reliability, and objective scoring (Aiken, 1987; Sax, 1989; Bennet and Ward, 1993). Moreover, the MC has gained great versatility in measuring objectives from the rote knowledge level to the most complex level (Sax, 1989). MC may also be used to assess macro-skills like listening for specific information, obtaining gist of what is being said, following directions, and following instructions (Hughes, 1989). However, the MC format has always been criticized for producing random guessing effect.

Hughes (1989, p. 60) states that “guessing may have a considerable but unknowable effect on test scores and as a result this has the effect of narrowing the range of scores.” Moreover, it is impossible to detect from the answer sheet the process through which the candidate respond correctly or wrongly. Furthermore, there is considerable doubt about their validity as measures of communicative language ability. Weir (1990, p. 44) states that “answering multiple-choice questions is an unreal task, as in real life one is rarely asked to choose from several options. Bachman and Palmer (1981) argue that tackling multiple-choice items in a test of LC may appear to be a test of the method rather than a test of the skill; that is, the format causes an undue effect on assessment of the trait (Weir, 1990). Harmful backwash effect is another disadvantage of the MC test because “there is a danger that practice for the test will have a harmful effect on learning and teaching” (Hughes 1989, p. 61).

Dictation: As a listening comprehension test, dictation is used in both languages teaching and testing. From the testing point of view, there is a significant conflict among experts regarding dictation as a technique of tapping LC. For instance, there are some difficulties for the examiners and examinees such as understanding the idiomatic expressions (Oller, 1979), and phrasal verbs (Davidson, 1989). Also there is no way to detect whether the examinees have understood the
meaning of the passage before (or even after) they have put it down on the paper. Many researchers in the field of testing (Taylor, 1981; Rost, 1990, James, 1984) maintain that “understanding spoken language is essentially an inferential process based on a perception of cues rather than straightforward matching of sound and meaning.” Lado (1961, p. 34, cited in Jafarpur and Yamini, 1993, p. 360) criticized dictation, calling it a test of auditory discrimination:

Since the order of words is given by the examiner as he reads the material, it does not test word order. Since the words are given by the examiner, it does not test vocabulary. It hardly tests aural perception of the examinee’s pronunciation; because the words can in many cases be identified by context of the students do not hear the sound correctly.

Davis and Rinvoluceri (1988) call dictation a well-worn and suspect technique and suggest abandoning such testing methods for less familiar ones.

Dicto-Comp: Dictation-Composition (Dicto-Comp) is an activity which involves dictation and composition. In Dicto-Comp, students are provided with auditory materials at a normal rate of speech, and then they are asked to write down from memory what they have heard. Dicto-comp combines text dictation and text reconstruction (Oller, 1979).

Dicto-Comp is essentially a memory-based task that seems to be almost an impossible undertaking to do for the ordinary mortals. This is because (a) “the capacity of short term memory is limited to about seven units, plus or minus two” (Miller, 1956), that is, “we can only hold in working memory about seven items at a time" (Nunan, 1991, p. 65); and (b) long term memory works with the propositional meaning of sentences, and not the actual words or syntactic devices that were used to utter the intended meaning. This is to say that the brain retains the information but discards the message, i.e., the exact words, phrases, or sentences that actually built the concept. Other disadvantages of Dicto-Comp are as follows: (1) Dicto-Comp includes not only listening comprehension but also processing, reconstructing, organizing and
writing materials. Therefore, the major goal of listening comprehension tests may be neglected. (2) Dicto-Comp scoring is more difficult than that of standard dictation. Verbatim scoring procedures are less applicable to this technique. "It presents many of the same scoring difficulties that will encounter in reference to essays in general". (Oller, 1979, p. 265) (3) Because of the complex nature of Dicto-Comp, this technique is not appropriate for the students at elementary and lower intermediate level of proficiency.

There is little research comparing task types for L2 listening comprehension. Brown, Anderson, Shadbolt, Lynch (1985) note that output tasks should not rely heavily, or exclusively, on memory or writing abilities. They also found that there is a great variation among L2 learners as to summarizing texts. Shohamy and Inbar (1991), working on the effect of question types on success in L2 listening tasks, found that test takers performed better on questions referring to local cues in the text than on those referring to global cues. They concluded that it is apparently more difficult to generalize, infer, and synthesize information than to look for data specific information. They also noted that low level test takers are more affected by the type of information demanded as opposed to high level examinees. They reported that most test takers who answered global questions were also able to answer local questions but not vice versa.

Therefore, the test method presented here represents an effort to move from global questions to local questions in context, from relying heavily to relying reasonably on memory, and from requiring exact responses to comprehensible and acceptable reproductions which is in accord with current methods of language teaching and testing.

The present study
This study intended to see what kinds of linguistic knowledge as well as listening strategies and cognitive processes are involved in taking Dicto-Phrase. According to Alderson, Clapham, and Wall (1999), an increasingly common aspect of test validation is to collect data on how
individuals respond to test items. An example of such learner-centered accounts is introspective information which is capable of revealing interesting insights into test performance (See Cohen1984, Faerch and Kasper 1987, and Grotjahn 1986). The simplest way to gather introspective data is retrospection. The disadvantage of retrospective data, however, is test takers may not remember why they answered items in a particular way. An alternative is to seek concurrent introspections, where the candidate thinks aloud, or performs an "online task" (Mackey and Gass, 2005) while responding.

Participants
Five female volunteer students in upper intermediate level of proficiency (as determined by Shiraz University Language Center Proficiency Test) studying English at Shiraz University Language Center were chosen. They were selected from a body of ten students who participated in a verbal protocol pilot study. The criteria for selecting the five language learner were their motivation to participate in the study and their ability to think aloud.

Instruments
The instrumentation of the study comprised two tests of D-P for the pilot study to familiarize test takers with the new procedure and a test of D-P for the main study. Also a semi-structured interview including retrospective questions was developed to gather further information after the think aloud procedure. (See the Appendix for the D-P used in the main study.)

Steps to D-P construction
In constructing D-P, (a) the idea units of the passage were determined; (b) after leaving the first (or the first few) idea unit(s) intact, a number of paraphrasable idea units (about half of the whole) which seemed to be more prominent were selected for the deletion purposes; and (c) relevant questions were developed for the missing idea units.
Scoring procedure of D-P
The criterion for scoring D-P was comprehensibility. Thus, to get credit, test takers needed to reproduce the intended propositions in their own words. The misspellings were ignored as long as the intended propositions were comprehensibly reproduced. However, if a change in the shape of the words affected the meaning of the intended proposition, test takers gained nothing.

The pilot study
In the pilot study, the students were briefed about the intended task. They were also familiarized with D-P test method and were trained on thinking aloud procedure. The instruction was given in their mother tongue, Persian, using the general think-aloud instruction drawn from Ericsson and Simon (1993, p. 378).

The main study
Each participant was tested separately. They were requested to give their verbal reports in their mother tongue as they engaged in completing the D-P tasks. Prompts like ‘keep talking’ or ‘talk aloud’ were used when the participant fell silent for a period of time. A liberal time was allowed since the act of generating a verbal report would slow down the performance. During the think-aloud, the researcher tried to stay outside the view of the participant to avoid intrusion. Following the think-aloud on the whole test, each student was asked retrospective questions. The questions were mostly intended to check for the strategies that students used to come up with answers. The verbal reports were tape-recorded for later transcription and analysis. On the whole, the investigator did every effort to create a stress-free atmosphere throughout data-collection.

The coding scheme
The purpose of developing a coding scheme is to capture commonalities among the behaviors of test takers in order to identify the possible common performance patterns (Green, 1998). The coding scheme was shaped by the aims of the study and by the task analysis.
A random sample of twenty percent of the protocols was examined to first develop the coding categories. The preliminary inspection suggested that the test takers attended to different sections of the item and made use of a range of distinct strategies. The students’ general categories of behavior were labeled (a) Response Strategies, and (b) Response Characteristics. The coding scheme is presented in Figures 1 and 2 below:
Twenty percent of protocols were coded twice by the researcher to find the degree of consistency in the two coding attempts. The intra-coder reliability was 0.95 for the protocols.
Results and Discussions
Response Strategies
Table 1 shows the test takers used more compensatory strategies than
cognitive. Within compensatory strategy subcategories, they utilized
more rereading strategy (57.4%). One interesting result may be the
significant difference between guessing intelligently (20.1%) and
guessing randomly (3.7%), the latter a characteristic of selected response
methods (e.g., the MC test method).

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Code</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Strategies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translation</td>
<td></td>
<td>41</td>
<td>18.8%</td>
</tr>
<tr>
<td>Compensatory Strategies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guessing intelligently</td>
<td></td>
<td>44</td>
<td>20.1%</td>
</tr>
<tr>
<td>Guessing</td>
<td></td>
<td>8</td>
<td>3.7%</td>
</tr>
<tr>
<td>Randomly</td>
<td></td>
<td>125</td>
<td>57.4%</td>
</tr>
<tr>
<td>Rereading strategy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>218</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2 displays the participants' attempts to search for the
intended clues in immediate text (45.5%), pretext (13.6%), or the post-
text (25%). On the whole, the data demonstrated that the test takers used
more linguistic knowledge (68.2%) than discourse knowledge (7%) to
fill in the blank spaces. Moreover, in their attempts to answer the
questions, the subjects sometimes activated their topical knowledge
(15.9%).


It is important to note that appealing to immediate text was more frequent than appealing to pretext and post-text. One could argue that the use of pretext and post-text use would increase if the passage on which D-P was made were more lexical and grammatical cohesive features. The same argument could be presented for less frequent use of discourse knowledge.

Table 3 below displays that the test takers made use of rereading strategy (78%) more than the revision (22%). By looking closer at Table 3, we recognize that reading the question was utilized less than reading the stem or response. One reason for higher frequency of rereading the stem can be the presence of more clues and information in the stem than the responses or the questions. In fact, the questions just provide hints to help test takers focus on the activity. In quite a few cases, the test takers attended to the questions in order to grammatically evaluate their responses. Moreover, with regard to revision, “first revision” was significantly more used than second, third, or fourth revisions.
Table 3
Rereading Strategy including reading different sections of D-P as well as revision

<table>
<thead>
<tr>
<th>Strategies</th>
<th>Frequency</th>
<th>Percentage</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rereading</td>
<td>Reading without change</td>
<td>42</td>
<td>33%</td>
</tr>
<tr>
<td></td>
<td>Reading the stem</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reading the response</td>
<td>38</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>Reading the question</td>
<td>19</td>
<td>15%</td>
</tr>
<tr>
<td>Revision</td>
<td>First Revision</td>
<td>17</td>
<td>13.4%</td>
</tr>
<tr>
<td></td>
<td>Second revision</td>
<td>5</td>
<td>3.9%</td>
</tr>
<tr>
<td></td>
<td>Third revision</td>
<td>4</td>
<td>3.1%</td>
</tr>
<tr>
<td></td>
<td>Fourth revision</td>
<td>2</td>
<td>1.6%</td>
</tr>
<tr>
<td>Total</td>
<td>127</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Response characteristics
Table 4 displays the frequencies of different response categories. With respect to the coding scheme, there are three different possible sufficient answers, namely, verbatim, acceptable (with to-the-point information), and comprehensible. In contrast, the insufficient responses included deviated answers, either with too little information, or too much (non-relevant) information, as well as inaccurate responses that consist of completely wrong information.

Table 4
The degree of accuracy: sufficient, deviated and inaccurate responses

<table>
<thead>
<tr>
<th>Response Categories</th>
<th>Frequency</th>
<th>Percentage</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sufficient</td>
<td>Verbatim</td>
<td>11</td>
<td>24.4%</td>
</tr>
<tr>
<td></td>
<td>Acceptable</td>
<td>13</td>
<td>28.9%</td>
</tr>
<tr>
<td></td>
<td>Comprehensible</td>
<td>7</td>
<td>15.6%</td>
</tr>
<tr>
<td>Deviated</td>
<td>Too Little</td>
<td>3</td>
<td>6.7%</td>
</tr>
<tr>
<td></td>
<td>Too much</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Inaccurate</td>
<td>11</td>
<td>24.4%</td>
<td>24.4%</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

The data demonstrated that 68.9% of the responses belong to the sufficient category. The deviated and inaccurate answers formed 6.7% and 24.4% of the whole data. Out of all the fifty questions only five contained no responses.
In the sufficient category, the acceptable and comprehensible responses almost doubled the verbatim ones. This is to say if the procedure recognized only exact replacements as correct, two thirds of the sufficient responses would be lost. This could be considered a major threat to the validity of the test results because different levels of knowledge could not be captured.

As Table 5 shows the data regarding time demonstrated that the participants invariably answered the questions and there were no significant differences between various subcategories of time (49.2% on the spot and 50.8% delayed responses).

<table>
<thead>
<tr>
<th>Table 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>The distribution of time use to answer the questions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Response Characteristics</th>
<th>On the Spot</th>
<th>Delayed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Before reading the question</td>
<td>17</td>
<td>27%</td>
</tr>
<tr>
<td></td>
<td>After reading the question</td>
<td>14</td>
<td>22.2%</td>
</tr>
<tr>
<td>Delayed</td>
<td>Remaining on the item</td>
<td>8</td>
<td>12.6%</td>
</tr>
<tr>
<td></td>
<td>Leaving the item</td>
<td>15</td>
<td>24%</td>
</tr>
<tr>
<td></td>
<td>Getting back to the item later</td>
<td>9</td>
<td>14.2%</td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

When the test takers had difficulty in answering the items on the spot, they resorted to different delay strategies, namely, remaining on the item, leaving the item, and getting back to the item later. Moreover, regarding the relationship between time and accuracy, as Table 6 shows,
83 percent of the "on-the-spot" answers were sufficient while the delayed but sufficient responses were 35 percent.

<table>
<thead>
<tr>
<th>Response characteristics</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>On the Spot</td>
</tr>
<tr>
<td>Sufficient</td>
<td>25 (83.3%)</td>
</tr>
<tr>
<td>Deviated</td>
<td>2 (6.7%)</td>
</tr>
<tr>
<td>Inaccurate</td>
<td>2 (6.7%)</td>
</tr>
<tr>
<td>No Response</td>
<td>1 (3.3%)</td>
</tr>
<tr>
<td>Total</td>
<td>30 (100%)</td>
</tr>
</tbody>
</table>

While performing on the test, two types of problems happened, namely, spelling and L₁–L₂ contradiction; that is, test takers comprehended the material but were not able to put it in L₂. The frequency of the problems was low, 3 and 7 instances, respectively.

In 5 cases, there were no responses. Of course, "no response" did not mean that the subjects did not do anything. The following processes and activities were observed in the "no response" category: (1) The sentence with the mutilation was repeatedly read. Each reading was followed by a long pause. This was to an effort to recall the answer or activate topical knowledge. Any attempt to respond was finally delayed. (2) The item was read in the context to see if more sentential clues could help recall the intended answer. (3) The intelligent guessing and rereading strategy were not successful; hence, they resorted to random guessing. (4) As a result of failure of all the above attempts, they left the item with no response.

**Conclusion**

In conclusion, the study revealed the following characteristics for the Dicto-Phrase method:
Requiring test takers to reproduce selected idea units makes D-P rather objective while allowing test takers to make acceptable and comprehensible (vs., exact) replacements makes it more subjective. Of course, subjectivity is the price for more valid as well as real-life results. We should remember that real-life use of language is not always exact.

D-P involves test takers in guessing intelligently rather than randomly. Moreover, most random guesses proved to be inaccurate. As a result, little room for chance score will remain that is a typical threat to the validity of selected-response testing procedures. Again, this seems to contribute to validity of the results produced by D-P due to diminishing measurement error.

The LTM seems to systematically forget certain linguistic features (lexical, phonological, and syntactic, and so on) but stores concepts and idea units which could be retrieved. In contrast, the STM seems to receive and hold the exact input or word for word chunks of language for a short time. Accordingly, the frequent reproduction of acceptable and comprehensible answers reveals that D-P diminishes the role of STM in making responses and boosts the role of LTM. This allows for comprehensible as well as acceptable responses for which examinees fail to replace the exact form. This, in turn, is expected to boost the validity of test scores in case the purpose of the measurement is comprehension rather than mere auditory memory.

D-P allows for different delay strategies, namely, remaining on the item, leaving the item, and getting back to the item later due to the presence of the backbone of the original passage to which examinees have already listened to. However, it seems that test takers have a better chance at deriving the right answer on the spot than when the response is delayed.

D-P allows for individual differences by its flexible scoring procedure and response time discussed in 1 and 4 above.
The following diagram summarizes all the processes that the subjects made use of in performing on the D-P test task.

**Pedagogical implications**

In addition to the features above, D-P seems to be an interface between teaching and testing which makes it an appropriate tool for tapping listening abilities for classroom purposes like listening for specific information and/or obtaining the gist of selected idea units.

Moreover, the highly contextualized nature of the test, due to including the foundations of the passage as well as the embedded questions, may contribute to the authenticity of the procedure. Finally, D-P appears to boost learners’ linguistic variation competence since it encourages the subjects to produce the same meaning, using a variety of forms of their own.

*Received 8 March, 2006*  
*Accepted 19 January, 2007*
References


The Construct Validation of Dicto-Phrase


Figure 3
The processes involved in performing on the D-P test
Appendix

The passage for Dicto-Phrase

Typical American teenagers are in fact very ordinary. They think their teachers make them work too hard; they love their parents but are sure they do not understand anything, and the most important things in their lives are their friendship. Most young people take jobs while they are in the school. They work at movie theaters, fast food restaurants, gas stations, and stores to pay for their entertainment. Maybe this is what makes them so independent of their parents at such a young age. It is not always easy to keep a job and do well in the school. But American children learn early that you have to work hard to win. Americans just like competitions, even in their time off. Extracted from Laird (1987, p. 8)

The Dicto-Phrase used in the main study

Typical American teenagers are in fact very ordinary. They think their teachers ........................................... (1. What do they think their teachers do?), they love their parents but are sure ......................... (2. What do they think of their parents?), and the most important things in their lives are ........................................... (3. What?). Most young people take jobs ....................................................... (4. When?), They work at movie theaters, .............................................................(5. Where else?), gas stations, and stores to .................................................. (6. Why do they work?). May be this is what makes them .................................................. (7. What?) at such a young age. It is not always easy to keep a job and .................................................. (8. What?). But American children learn early that .................................................. (9. What do they learn?). American just like competitions, even .................................................. (10. When?)
The Construct Validation of Dicto-Phrase