Matching test (MT) construction techniques were compared. These included: (a) MT instruction that limit the number of times responses can be selected to once, rather than more than once; (2) MTs that are organized in groups of five premises, or stems, rather than in groups of ten premises; and (3) MTs designed to measure knowledge, rather than a higher-order achievement level such as comprehension or synthesis. The subjects were 196 undergraduate students enrolled in an introductory educational psychology course offered at the University of Pittsburgh. Significantly higher scores were observed for MTs organized in groups of five premises and for MTs designed to measure higher-order achievement levels. A recommendation is made to organize MTs in groups of five premises. The absence of significant interactions suggests that this recommendation may apply for MTs designed to measure either knowledge or some higher order achievement. Also, a special advantage of MTs in the assessment of partial knowledge is discussed. (Author/RC)
THE CONSTRUCTION OF MATCHING TESTS: AN EMPirical STATEMENT

Gregory A. Shannon
Pennsylvania Department of Education

Presented at the Annual Meeting of the National Council on Measurement in Education

Washington, D.C.
April 1975
The Construction of Matching Tests: An Empirical Statement

Gregory A. Shannon
Pennsylvania Department of Education

A matching test (MT) is a form of a multiple-choice test for which the premises, or stems, and the response options are each typically arranged in separate columns. The testees are asked to match each premise with the appropriate response or responses. An MT is usually broken down into groups of premises and responses which are called matching exercises (MEs). Most discussions of MT construction were included in publications for which empirical support was not apparent (Odell, 1928, Lang, 1930, Ebel, 1951, and Wesman, 1971).

Constructional aspects of major concern include MT instructions, ME length, and cognitive achievement levels measurable by MTs. As opposed to restricting the number of times a given response option may be selected, Wesman (1971) felt that non-restricted MT instructions would yield the best type of MT scores. What Wesman meant by best MTs seemed to be MTs which would be more difficult to complete, and thus, less susceptible to testee traits such as guessing and risk taking abilities.

Suggested ME lengths have ranged from five premises (Ebel, 1951) to thirty or more (Good, 1927). ME length was defined by the number of premises assigned to the ME. Thus, an ME which

This paper is a summary of the author's Master's thesis at the University of Pittsburgh.

I would like to thank Dres. Henry Hausdorff, Richard Cox, Robert Guthrie, Stanley Jacobs, and Charles Stegman for serving on my committee.

3
consisted of five premises was defined as a five-premise ME. In general, it was agreed that MTs would be more difficult to complete as ME length is increased.

Although MTs may be constructed to measure knowledge (Wassman, 1971) and higher-order achievement levels (Fay, 1929, Ebel, 1951, and Secord, 1952), the effect of measuring lower-order and higher-order cognitive achievement levels (Bloom, 1956) upon MT scores remains to be examined. It does seem reasonable that MTs constructed to measure higher-order achievement levels would be more difficult than MTs constructed to measure knowledge.

It has been the author's experience that MTs are commonly being constructed and used by classroom teachers to assess achievement. Since guidelines employed by teachers in constructing MTs are very much in need of research fortification, it is the purpose of this study to empirically strengthen this support. The MT construction techniques investigated in this study included:

(a) the use of MT instructions that ask the testee to select a given response option only once, as opposed to, MT instructions that permit the testee to select a given response option once or more, (b) constructing MTs that consisted of five-premise MEs, as opposed to, ten-premise MEs, and (c) constructing MTs as measures of knowledge, as opposed to, constructing MTs as measures of higher-order achievement levels such as comprehension or synthesis.

It was hypothesized that higher MT scores would be observed when the MT instructions restrict response selection to once, when the MTs consist of five-premise MEs, and when the MTs are constructed as measures of knowledge. It was also hypothesized that no
interactions would be observed between the constructional variables under investigation.

METHOD

Materials

A premise-response (P-R) pool was constructed to conform to the final instructional phase of an introductory educational psychology course at the University of Pittsburgh. The resources consisted of the course texts and an existing pool of multiple-choice test items. Premises were constructed and paired with correct responses. The initial pool of P-R pairs was reviewed by the course instructor for content validation purposes and revised. Examinations were also made to insure that each premise listed had been assigned a unique response. For the purpose of classifying the P-R pairs into achievement levels, three graduate students who had completed introductory educational psychology courses served as judges. They were each presented with the revised P-R pool along with written instructions which asked them to select the P-R pairs which they judged to be measures of knowledge as defined by Krathwohl and Payne (1971); P-R pairs that were unanimously judged to be measures of knowledge were classified accordingly and P-R pairs unanimously judged to be measures of some other cognitive ability were classified as measures of some higher-order achievement level. It was assumed that knowledge occupied the lowest cognitive level of the Taxonomy defined by Bloom and that the hierarchy was exhaustive. P-R pairs which only two of the judges classified were reviewed by a professor within the Department of Educational Psychology.
and either classified or eliminated.

The remaining P-R pairs were constructed into MTs and reviewed by the course instructor and two graduate assistants to insure that each response originally paired with a given premise would remain the most reasonable choice. A final revision was made. All MTs consisted of thirty premises and thirty-six responses.

The MT instructions consisted of two forms; restricted and unrestricted. Restricted instructions were stated as follows:

Complete each statement (numbered) with the most correct response (lettered). Enter each response in the blank preceding each statement. Do not select any given response more than once. Please do not guess.

The unrestricted instructions were stated as follows:

Complete each statement (numbered) with the most correct response (lettered). Enter all responses in the blank preceding each statement. Any response may be selected more than once. Please do not guess.

The two ME lengths were five-premise MEs (with six responses) and ten-premise MEs (with twelve responses). Each MT form consisted of thirty premises broken down into either five-premise MEs or ten-premise MEs.

The three classifications of cognitive achievement which the MTs were constructed to measure were lower-order, higher-order, and a combination of lower-order and higher-order achievement.

Subjects

The Ss were one-hundred and ninety-six undergraduate students enrolled in an introductory educational psychology course offered by the Department of Educational Psychology at the University of
Pittsburgh. The majority of the Ss were female. The MTs were administered as pre-tests for the final instructional unit of the course. The scores were not included in the Ss' course grades.

Procedure

The Ss were assembled in a small auditorium. Twelve forms of the MTs (two forms of MT instructions X two ME lengths X three levels of cognitive achievement) were randomly assigned to the Ss. The Ss were advised that certain of the MTs were different in regard to the instructions and the ME lengths. Testing was completed within the hour normally scheduled for the course. Because of typographical errors one premise from six of the MT forms and three premises from two of the MT forms were not included in the data analysis.

Design

A 2 X 2 X 3 full rank design was employed. The data were analyzed using the univariate procedure described by Timm and Carlson (1973). Achievement was the primary dependent variable. Although subtest scores were computed for grading purposes, only the total scores were included in the data analysis.

RESULTS

Percentage scores were computed for the MTs. Means, standard deviations, and ranges (by treatment groups) are included in Table 1.
<table>
<thead>
<tr>
<th>Cell (Group)</th>
<th>N</th>
<th>Number of Items</th>
<th>Mean</th>
<th>Standard Deviation&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC 111</td>
<td>4</td>
<td>29</td>
<td>81.90</td>
<td>7.11</td>
<td>75.86 - 89.66</td>
</tr>
<tr>
<td>ABC 211</td>
<td>19</td>
<td>29</td>
<td>68.97</td>
<td>27.34</td>
<td>6.90 - 96.55</td>
</tr>
<tr>
<td>ABC 121</td>
<td>17</td>
<td>29</td>
<td>65.31</td>
<td>10.73</td>
<td>48.28 - 86.21</td>
</tr>
<tr>
<td>ABC 221</td>
<td>10</td>
<td>29</td>
<td>62.07</td>
<td>19.71</td>
<td>27.59 - 89.66</td>
</tr>
<tr>
<td>ABC 112</td>
<td>25</td>
<td>27</td>
<td>79.26</td>
<td>10.69</td>
<td>44.44 -100.00</td>
</tr>
<tr>
<td>ABC 212</td>
<td>17</td>
<td>27</td>
<td>76.69</td>
<td>10.04</td>
<td>59.26 - 92.59</td>
</tr>
<tr>
<td>ABC 122</td>
<td>18</td>
<td>29</td>
<td>67.43</td>
<td>13.34</td>
<td>41.58 - 89.66</td>
</tr>
<tr>
<td>ABC 222</td>
<td>10</td>
<td>29</td>
<td>71.03</td>
<td>11.86</td>
<td>51.72 - 89.66</td>
</tr>
<tr>
<td>ABC 113</td>
<td>14</td>
<td>30</td>
<td>84.05</td>
<td>7.53</td>
<td>70.00 - 96.67</td>
</tr>
<tr>
<td>ABC 213</td>
<td>31</td>
<td>30</td>
<td>83.23</td>
<td>11.59</td>
<td>40.00 -100.00</td>
</tr>
<tr>
<td>ABC 123</td>
<td>12</td>
<td>30</td>
<td>81.11</td>
<td>7.70</td>
<td>70.00 - 93.33</td>
</tr>
<tr>
<td>ABC 223</td>
<td>19</td>
<td>30</td>
<td>73.68</td>
<td>15.63</td>
<td>20.00 - 93.33</td>
</tr>
</tbody>
</table>

<sup>a</sup>These unbiased estimates were not adjusted. Adjustments may be made by multiplying them by \((100/N^2)\), with \(N\) being the appropriate number of items.
The F-ratios computed for both the B factor (ME length) and the C factor (achievement levels) were significant well beyond the .01 level. The analysis of variance summary table is presented below (Table 2). The means observed for factor B were 78.91 and 70.10, respectively. The mean observed for five-premise MEs was higher than the mean observed for ten-premise MEs.

**TABLE 2**
Results of ANOVA for Relationship among MT Instructions, ME Length, Achievement Levels and Achievement

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>F-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT Instructions (A)</td>
<td>1</td>
<td>386.63</td>
<td>2.20</td>
</tr>
<tr>
<td>ME Length (B)</td>
<td>1</td>
<td>3473.70</td>
<td>19.76*</td>
</tr>
<tr>
<td>Achievement Levels (C)</td>
<td>2</td>
<td>2724.71</td>
<td>7.75*</td>
</tr>
<tr>
<td>A X B</td>
<td>1</td>
<td>26.99</td>
<td>.15</td>
</tr>
<tr>
<td>A X C</td>
<td>2</td>
<td>290.81</td>
<td>.83</td>
</tr>
<tr>
<td>B X C</td>
<td>2</td>
<td>328.53</td>
<td>.93</td>
</tr>
<tr>
<td>A X B X C</td>
<td>2</td>
<td>390.36</td>
<td>1.11</td>
</tr>
<tr>
<td>Within</td>
<td>184</td>
<td>32340.67</td>
<td>175.76</td>
</tr>
</tbody>
</table>

*p<.01

The means observed for factor C were 67.38 for lower-level achievement, 74.42 for composite, and 80.66 for higher-level achievement. A systematic increase in scores across achievement levels was indicated.

The F-ratio computed for the BC interaction served also to suggest the absence of a significant interaction. Thus, inter-
pretation of the main effects, B and C, was simplified. The systematic increase of means observed for factor C (achievement levels) was observed for two levels of B (ME Length). This increase was in the order of lower-level, composite, and higher-level achievement, respectively. Expressed in terms of factor B, means observed for five-premise MEs were higher than the means observed for ten-premise MEs, across the three levels of factor C.

DISCUSSION

The findings of the present investigation indicate the following conclusions:

1. The experimental hypothesis that unrestricted MT instructions would result in lower mean scores than restricted instructions was not supported.

2. The experimental hypothesis that ten-premise MEs would result in lower mean scores than five-premise MEs was supported.

3. The experimental hypothesis that mean scores would decrease as achievement level was increased was not supported, although the reverse situation was supported.

4. The experimental hypothesis that there would be no significant interactions among the three main factors was supported.

From the findings it is recommended that the number of premises to be included within a matching exercise be held to approximately five premises, with a greater number of response options to reduce guessing. Increasing matching exercise length places greater demands on testees' skills which may not be of immediate concern such as reading comprehension, attentiveness, and organization. The practice of test toughening by increasing the length of the
matching exercises will probably reduce the reliability and interpretability of the scores. The absence of significant interactions suggests that the above recommendation may reasonably apply for matching tests designed to measure either knowledge or some higher-order achievement level such as comprehension or synthesis. This is good because most matching tests constructed for classroom purposes measure a mixture of lower-order and higher-order achievement.

Also, it is suggested that matching tests have an advantage over multiple-choice tests in the assessment of partial knowledge. While constructing the matching tests, the author had an opportunity to make subjective comparisons with the multiple-choice test items based upon the same material. Multiple-choice tests consist of items which include one stem and a specified number of responses. Matching tests consist of matching exercises which include a response for each premise. The operation of partial knowledge is present with both types of tests. However, a four-option multiple-choice test item would normally fail to assess testees' knowledge of three of the four options. Whereas, a four-premise matching exercise (with five response options) may fail to assess knowledge of only one distractor. It seems reasonable that testees would have a greater opportunity to demonstrate their total knowledge on matching tests rather than on multiple-choice tests.

The matching tests became easier as the achievement levels were increased. The explanation may follow from the fact that higher institutions prepare students to become the thinkers, rather
than the memorizers of society. In time, students learn to adjust to this fact. The author feels that this is the major reason why the students who took the matching tests were better prepared to demonstrate their understanding of the course subject matter, rather than to demonstrate their ability to recall details.

REFERENCES


