DEVELOPMENT OF A MACHINE TRANSCRIPTION
WORK SAMPLE TEST
FOR SECRETARIAL SELECTION

THESIS

Presented to the Graduate Council of the
North Texas State University in Partial
Fulfillment of the Requirements

For the Degree of

MASTER OF ARTS

By

Deborah Frances Kaye, B. A.
Denton, Texas
August, 1981

The study described the development of a standardized, normed, content-valid machine transcription test which could be used to evaluate the ability of secretarial applicants to type a mailable copy of a business letter from a dictated tape recording. The test was based on a thorough job analysis and was pretested using a pilot study with job incumbents to confirm its feasibility. Normative data were developed from 50 job applicants. Interrater reliability was statistically significant ($r = .85, p < .05$). The test was adopted for use at the headquarters office of a major oil and gas producing company.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF ILLUSTRATIONS</td>
<td>v</td>
</tr>
<tr>
<td>Thesis</td>
<td></td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Validation Strategy</td>
<td></td>
</tr>
<tr>
<td>Behavioral Consistency Model</td>
<td></td>
</tr>
<tr>
<td>Content Validity</td>
<td></td>
</tr>
<tr>
<td>Test Construction</td>
<td></td>
</tr>
<tr>
<td>Utility</td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>10</td>
</tr>
<tr>
<td>Job Analysis</td>
<td></td>
</tr>
<tr>
<td>Review of job descriptions</td>
<td></td>
</tr>
<tr>
<td>Interviews with incumbents</td>
<td></td>
</tr>
<tr>
<td>Work samples</td>
<td></td>
</tr>
<tr>
<td>Test Development</td>
<td></td>
</tr>
<tr>
<td>Pilot Study</td>
<td></td>
</tr>
<tr>
<td>Norm Development</td>
<td></td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>15</td>
</tr>
<tr>
<td>Appendix</td>
<td>20</td>
</tr>
<tr>
<td>References</td>
<td>22</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Typical Characteristics of Documents Transcribed by Machine</td>
<td>11</td>
</tr>
<tr>
<td>2. Punctuation and Style Characteristics in Work Samples</td>
<td>12</td>
</tr>
</tbody>
</table>
LIST OF ILLUSTRATIONS

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Frequency Distribution of Total Errors for Normative Sample</td>
<td>15</td>
</tr>
<tr>
<td>2.</td>
<td>Scattergram of Total Errors by Time to Complete Test for Normative Sample</td>
<td>16</td>
</tr>
<tr>
<td>3.</td>
<td>Scattergram of Total Errors by Key Errors for Normative Sample</td>
<td>18</td>
</tr>
</tbody>
</table>
DEVELOPMENT OF A MACHINE TRANSCRIPTION WORK SAMPLE TEST FOR SECRETARIAL SELECTION

The principles set forth in the Uniform Guidelines on Employee Selection Procedures, henceforth Guidelines (1978), were that employer policies or practices which discriminated against any race, sex, or ethnic group were illegal under federal law unless justified by business necessity. The Guidelines charged the users of selection procedures with the responsibility for validating those procedures. Validity is the degree to which inferences from scores on tests or assessments are justified or supported by evidence (Principles for the Validation and Use of Personnel Selection Procedures, henceforth Principles, 1980). When the classical validity model is applied in personnel selection, tests are typically used as signs of predispositions to behave in certain ways on the job. Use of this model has not ameliorated many of the difficulties encountered in predicting job performance, such as the low validities obtained for selection procedures.

Validation Strategy

As a solution to these difficulties, Wernimont and Campbell (1968) proposed that the classical model be altered and introduced the concept of behavioral consistency. Behavioral consistency means that "The best indicator of future performance is past performance" (p. 372). In the application
of behavioral consistency to selection, meaningful samples of work behavior, such as work sample tests, are used as predictors of future work behavior. Additionally, the authors suggested that making test content more relevant to work would have the advantages of diminishing the problems of "faking" and response sets and reducing charges of discrimination and invasion of privacy in testing. The present study was developed to describe the methods used in the construction of a work sample test designed in accordance with the accepted psychological and legal principles for employee selection.

**Behavioral Consistency**

Work samples are classified into two types—motor, if the task involves the physical manipulation of things, and verbal, if the problem situation is primarily language-oriented or people-oriented. In a review of work sample tests which had been designed for specific tasks, Asher and Sciarrino (1974) concluded that complex tasks such as work samples were likely to be more powerful predictors than tests of single dimensions because work samples had more in common with the criteria. Motor work samples had the most visible connection with the criterion of job proficiency and higher validity coefficients than verbal work samples and tests of intelligence, mechanical aptitude, finger dexterity, personality, and spatial relations.
The practical utility of motor work samples has been demonstrated in the selection of maintenance mechanics (Campion, 1972). Verbal work samples (specifically simulation exercises in assessment centers) have been especially successful in demonstrating the advantages of the consistency approach in selecting managers and salesmen (Bray and Campbell, 1968; Hinrichs, 1969).

In a laboratory study of the predictive validity of a work sample, Mount, Muchinsky, and Hanser (1977) found both the predictive and concurrent validity coefficients obtained for work samples to be comparable to coefficients obtained for paper-and-pencil tests. Work samples were substantively similar to the criterion, and the amount of shared variance between work sample and criterion was higher than the amount for paper-and-pencil tests in each group. This finding was consistent with the concept of Asher and Sciarrino (1974) that a predictor which most closely resembles the criterion task should account for greater variance than a measure without a point-to-point correspondence with the criterion. Mount et al. (1977) concluded that an additional advantage of work samples was that they made test content more relevant to the work setting, a vital consideration in complying with the spirit of the Guidelines (1978).

Content Validity

In order to meet the standards set by the Guidelines (1978), an employer must be able to demonstrate the job
relatedness of a selection procedure. There are three concepts which have been used for this purpose: criterion-related validity, construct validity, and content validity. Content validity is a relationship between job performance and a test that is self-evident because the test includes a representative sample of job tasks (Principles, 1980).

A work sample test could be supported by a content validity argument to the extent that it had a representative sampling of job content and measured only those knowledges, skills, or abilities which were a prerequisite to employment. Skills or abilities which were to be measured should have been operationally defined in terms of observable aspects of work behavior and work products.

Recent literature has suggested that what was known as content validity was actually either content-oriented test construction or a special case of construct validity (Guion, 1974, 1977, 1978a, 1978b; Tenopyr, 1977). Ebel (1977) has stated that if applications of content validity did not involve inferential validity, then content validity should have been more properly called content reliability or job sample reliability.

Psychologists have addressed these differing views. Noting that correlations between essentially the same behaviors were sometimes labeled as reliabilities instead of as validities, Mount et al. (1977) maintained that if the purpose of a selection program was to obtain predictor
measures which were as similar to the criteria as possible, then the term **validity** should have been applied to correlations obtained between a predictor and a criterion even when those two measures were conceptually similar.

In an effort to sharpen the content validity concept, Lawshe (1975) defined content validity as "the extent to which communality or overlap exists between (a) performance on the test under investigation and (b) ability to function in the defined job performance domain" (p. 566). The nature of job performance domains comprised a continuum from behavior that was highly abstract to behavior that was directly observable. Lawshe further specified that as the behavior elicited by the test more closely approximated a true work sample of the job performance domain, people who knew the job were more competent to assess the content validity of the test. He concluded that

content validity analysis procedures are appropriate only when the behavior under scrutiny in the job performance domain falls at or near the 'observation' end of the continuum. . . . such analyses are essentially restricted to (1) simple proficiency tests, (2) job knowledge tests, and (3) work sample tests. (p. 566)

**Test Construction**

The technical standards for demonstrating content validity outlined in the Guidelines (1978) required that a job analysis be conducted to determine the work behaviors
necessary for successful performance. Recently, researchers have highlighted the importance of conducting thorough job analyses in order to identify valid predictors and meaningful job effectiveness criteria. Cornelius, Carron, and Collins (1979) examined the effects of three job analysis models (task-, worker-, and ability-oriented) on subsequent job classification. In a task-oriented analysis which focused on actual work activities, jobs were divided into elemental units. The authors suggested that the task approach was useful in developing selection systems for jobs in which employees were expected to demonstrate task performance without training (e.g., typists). Work sample and situational tests that emphasized task performance were especially appropriate in this case. Prien (1977) recommended that content validation strategies utilize a combination of task- and worker-oriented methods.

A comparison of a job knowledge test constructed on the basis of a systematic job analysis with a commercial employment test was conducted by Kesselman and Lopez (1979) to determine the fairness of each test for minority and non-minority applicants for accounting jobs. Significant validity coefficients for a wide range of job success criteria had been reported for the commercial test. The authors hypothesized that predictors chosen without job analysis information would not measure the most important determinants of job success, and might measure factors that were not
related to acceptable job performance. Results indicated that the job knowledge test was a valid and unbiased predictor of job performance criteria. The commercial test produced adverse impact and lacked validity. These findings supported the contribution of job-analyses-based selection procedures to the mitigation of test bias problems.

Guion (1978b) proposed that the method of test construction would, without further empirical investigation, provide a sufficient basis for considering a test an acceptable job-related measure for employment purposes if the following six conditions were met:

1. The test content domain should be defined in terms of behavioral content with a generally accepted meaning.

2. The test content domain should be defined unambiguously.

3. The test content domain should be relevant to the purposes of testing.

4. Qualified judges must agree that the test content domain has been adequately sampled.

5. Responses to the content sample must be reliably observed and evaluated.

6. Opportunities for contamination in the evaluation of responses should be slight. (p. 502)

Schoenfeldt, Schoenfeldt, Acker, and Perlson (1976) successfully implemented a content-validated test of
industrial reading ability in the screening of applicants for entry-level positions in a large company. They viewed content-oriented test development or content validation as one answer to the recent legislative mandate for extensive reexamination of recruitment and selection procedures.

Utility

Prior to the implementation of a valid selection procedure, judgments should have been made about the relative costs and benefits of the procedure to the organization and its employees. The costs of testing include the direct cost of testing each applicant as well as the costs of developing, validating, and gaining public acceptance for a selection procedure (Cronbach, 1980). Enhancing the selection of qualified employees through the use of valid selection procedures has been the primary benefit of employment testing.

Standards for selection must have been established before a test could be used for hiring purposes. Principles (1980) addressed this issue.

Selection standards may be set as high or low as the purposes of the organization require, if they are based on valid predictors. This implies that (a) the purposes of selection are clear and (b) they are acceptable in the social and legal context in which the employing organization functions. In usual circumstances, the relationship between a predictor and a criterion may be assumed to be linear. Consequently,
selecting from the top scores on down is almost always the most beneficial procedure from the standpoint of the organization if there is an appropriate amount of variance in the predictor. Selection techniques developed by content-oriented procedures and discriminating adequately within the range of interest can be assumed to have a linear relationship to job behavior. (p. 18)

The placement of critical scores should maximize the benefit of an employment procedure to an organization through the selection of the most qualified applicants. In this regard, *Principles* (1980) has stated:

The only justification which can be demanded is that critical scores are determined on the basis of a reasonable rationale. This may involve such factors as estimated cost-benefit ratio, selection ratio, success ratio, social policies of the organization, or judgments as to required knowledge, skill, or ability on the job. (p. 19)

A similar view has been expressed in the *Guidelines* (1978). Where cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force. (p. 38298)

The present study was developed to describe the methods used in the development of a content-validated test of machine
transcription skill found to be necessary for entry-level secretarial positions in selected departments within a major oil and gas producing company. The skill of transcribing by machine entailed listening to a dictated tape recording of a business letter and typing it in mailable copy. No commercially published standardized test has been available for this purpose. All technical standards for content validity studies from the Guidelines (1978) were followed during test development.

Method

Job Analysis

Review of job descriptions. From an examination of position analyses for entry-level secretarial jobs, it was determined that machine transcription was a required duty in one administrative and two accounting departments. Specifically, the principal duties of the position included the typing of routine transmittals, reports, statements, requisitions, or other correspondence from recorded dictation.

Interviews with incumbents. Interviews were held with two job incumbents in each of the three departments to review the following characteristics of documents transcribed by machine:

(a) document content (general business or technical)
(b) document length in pages
(c) document type (internal company correspondence or outgoing mail)
(d) time required to produce a mailable document
(e) form of document typed directly from tape
(rough draft or final/mailable copy)
(f) speed of dictation (slower than, equal to, or
faster than conversation)

Results are presented in Table 1.

Table 1

Typical Characteristics of Documents Transcribed by Machine

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Administrative</th>
<th>Accounting 1</th>
<th>Accounting 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>content</td>
<td>general</td>
<td>general</td>
<td>general</td>
</tr>
<tr>
<td>length</td>
<td>1-2 pp.</td>
<td>2 pp.</td>
<td>2-3 pp.</td>
</tr>
<tr>
<td>type</td>
<td>internal/</td>
<td>internal/</td>
<td>internal/</td>
</tr>
<tr>
<td></td>
<td>outgoing</td>
<td>outgoing</td>
<td>outgoing</td>
</tr>
<tr>
<td>time</td>
<td>10-20 min.</td>
<td>5-25 min.</td>
<td>20-40 min.</td>
</tr>
<tr>
<td>form</td>
<td>rough/final</td>
<td>rough/final</td>
<td>rough</td>
</tr>
<tr>
<td>speed</td>
<td>slow</td>
<td>slow</td>
<td>slow/fast</td>
</tr>
</tbody>
</table>

All incumbents stated that they used an IBM Selectric
typewriter or its equivalent and a transcribing machine
with a footpedal and earphones. Materials for correcting
typing errors and a dictionary were always accessible. It
was sometimes necessary to consult the dictator of the tape
to clarify words or instructions. Typing accuracy and speed
of completion were both important factors in supervisors' assessments of secretarial performance on transcription tasks.

Work samples. Sample documents were collected from the selected departments and examined for similarities in punctuation and style characteristics. The typical document was a two-page internal or outgoing letter of general business content. All company correspondence was produced according to the uniformly recognized specifications of full block style. Table 2 summarizes the most frequently occurring punctuation and style characteristics.

Table 2
Punctuation and Style Characteristics in Work Samples

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Total Number* of Occurrences of Characteristic</th>
<th>Percentage of Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>All capital letters (e.g., APA)</td>
<td>180</td>
<td>25</td>
</tr>
<tr>
<td>Dates in text</td>
<td>110</td>
<td>15</td>
</tr>
<tr>
<td>Numbers</td>
<td>105</td>
<td>15</td>
</tr>
<tr>
<td>Hyphenated words</td>
<td>82</td>
<td>11</td>
</tr>
<tr>
<td>Underlined words</td>
<td>76</td>
<td>10</td>
</tr>
<tr>
<td>Parentheses</td>
<td>49</td>
<td>7</td>
</tr>
<tr>
<td>Percent and dollar signs</td>
<td>31</td>
<td>4</td>
</tr>
<tr>
<td>Double quotation marks</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Slashes</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>Colons and semi-colons</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>3-column charts</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Carbon copy notations</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Enclosure and attachment notations</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>717</td>
<td>99</td>
</tr>
</tbody>
</table>

* 41 Documents
Test Development

Utilizing information obtained from the job analysis, a test letter was written. It was representative of the typical characteristics of a document which was transcribed by machine. The letter was two pages long and contained general content. An outgoing correspondence format was used since the internal correspondence format required knowledge of certain company-specific style conventions (beyond full block style) which an applicant could not be expected to bring to the job. See Appendix A for the complete test directions.

The letter was recorded at slower than conversational speed and the tape played for approximately four minutes if uninterrupted. Since applicants were unable to question the dictator during the test, most punctuation and style characteristics were dictated on the tape.

Criteria for evaluation were based on the usual and customary standards for evaluating business correspondence. These included correctly following all written and taped directions. Scoring instructions for raters detailed the method for checking items. Each incorrect item (e.g., a misspelled word, omitted punctuation) was counted as one error.

Pilot Study

A group of twelve job incumbents took the test under simulated working conditions. Their performance was
evaluated by two raters based on the scoring criteria. Mean number of errors determined by Raters 1 and 2 were 5.75 (SD = 3.91) and 5.66 (SD = 3.67), respectively. The interrater reliability was high and statistically significant (r = .92, p < .05). The mean and standard deviation of the number of minutes taken to complete the test (time) were 24.16 and 7.96. (This mean time of 24.16 minutes approximated the 20-minute average of the times reported by job incumbents in interviews.) There was a moderate relationship between time and number of errors (based on the scores of Rater 1, r = .52, p < .05).

Incumbents were asked to give their expert opinion of the degree to which the test sampled the skill needed for successful machine transcription on the job. They reported that all aspects of the task of transcribing the test letter were similar to the on-the-job transcription task for each of the characteristics determined in the job analysis. The feasibility of the machine transcription test was supported by the pilot study.

**Norm Development**

Normative data were collected on a sequential 100% sample of 50 applicants for entry-level secretarial positions at the headquarters office of a major oil and gas producing company. An analysis of the scores from this group was used to establish test norms. Interrater reliability was obtained to ensure that the scoring instructions allowed for consistent evaluation of applicants.
Results and Discussion

The mean number of errors determined by Rater 1 and Rater 2 are 10.92 (SD = 7.93) and 9.36 (SD = 7.37), respectively. Interrater reliability is high and statistically significant ($r = .85, p < .05$). The mean and standard deviation of the number of minutes taken to complete the test are 59.90 and 23.97. The correlation of time with number of errors (based on Rater 1) is not statistically significant ($r = .17, p > .05$). Figure 1 shows the distribution of errors for the normative sample.

![Graph showing distribution of errors for the normative sample.]

Figure 1. Frequency Distribution of Total Errors for Normative Sample

Based on the distribution of errors in the normative group, it is determined that applicants making more than 13 errors do not meet the normal expectations of acceptable proficiency within the work force (Guidelines, 1978). In accordance with the company's goal of selecting the most qualified applicants, the critical score is set at 13.
Based on this critical score, 78% (11) of the applicants would be accepted and 22% (11) would be rejected.

The amount of time taken to complete the test ranged from 30 to 135 minutes (M = 59.90). The relationship of time to errors is presented in Figure 2.

Figure 2. Scattergram of Total Errors by Time to Complete Test for Normative Sample.

Of the applicants who meet the critical score, 22% (11) exceed the mean time for completion of the test. However, four of these 11 applicants exceed the mean
time by not more than 15 minutes. Company selection practices concerned with the costs of testing required that a reasonable time limit be set. With a one hour time limit, 72% (28) of the applicants who meet the critical score for errors would be accepted. It was assumed that with advance knowledge of the time limit, applicants who would otherwise exceed it by not more than 15 minutes might be able to pace themselves to complete the test in one hour with few errors. The time limit was therefore set at one hour.

The level of detail in the machine transcription test resulted in lengthy scoring times. To reduce the time demands placed on administrative clerks who had been trained to score the test, an abbreviated scoring system was developed. From the sample of 50 test letters, a count was made of the frequency with which errors occurred on each item (e.g., each word, punctuation mark, or phrase). The 20 items on which errors were most frequently committed were noted, and it was hypothesized that the number of these errors which occurred in any given test would have a meaningful relationship with the total number of errors in that test. The score of each test was then recalculated as the number of the twenty most frequent errors it contained. Figure 3 shows the relationship between these errors and total errors. The part/whole correlation of most frequent errors with total errors is high and statistically
significant \((r = .88, p < .05)\). The 20 most frequent errors are designated "key errors."

![Scattergram of Total Errors by Key Errors for Normative Sample](image)

Figure 3. Scattergram of Total Errors by Key Errors for Normative Sample.

As Figure 3 illustrates, all applicants who meet the critical score of 13 also have nine or fewer key errors; no applicant who meets the critical score of 13 total errors would be rejected on the basis of the key scoring system. The final scoring procedure entails locating each key item in the test letter and comparing it to a correct listing of
the item. The resulting test evaluation procedure combined the key scoring system with the one-hour time limit and would distinguish those applicants who were most qualified in the tested skill from a large applicant pool. Most importantly, the test would identify the applicants who met the standards of ability required by the employing company. Specifically, the selected applicants would excel on the two machine transcription skill factors deemed important by supervisors--typing accuracy and speed of completion. Use of the test for selection purposes provides maximum benefits to the employing organization through the selection of the most qualified applicants, and to the applicants by allowing them to directly demonstrate their machine transcription skill under conditions that simulate an on-the-job situation.

The present study described the development of a standardized, normed, content-valid machine transcription test which can be used to evaluate the ability of secretarial applicants to type a mailable copy of a business letter from a dictated tape recording. It was developed and validated according to the appropriate federal guidelines and psychological principles. This machine transcription test is currently in use at the headquarters office of the company in which this study was conducted.
Appendix A

Machine Transcription Test Directions

(today's date)

Secretarial Applicant
Company Name
Company Address
City, State
Zip Code

Dear Secretarial Applicant:

This is a test of your ability to produce a MAILABLE COPY of a business letter from a dictated tape recording.

Turn in only a mailable copy of the letter. You may type a rough draft from the tape first, if you wish. The letter is 2 pages long. The margins are already set for a 65-space line. Place the date 15 lines from the top of the page.

Type the mailable copy of the letter in FULL BLOCK STYLE. As an example for you to use, these directions are typed in full block style. You will be required to type a CHART in the body of the letter. The chart you are to use is on the back of these directions.

Use today's date for the letter. Use your name to close the letter. You may use correction fluid to correct typing errors. You may use the dictionary to look up the spelling of words. You may start, stop, and reverse the transcribing machine as you wish.

When you are finished, please give your mailable copy to the test administrator immediately.

Sincerely,

Test Administrator
<table>
<thead>
<tr>
<th>SEMINAR</th>
<th>COST</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science/Technology</td>
<td>$250</td>
<td>New York</td>
</tr>
<tr>
<td>Education</td>
<td>$225</td>
<td>Houston</td>
</tr>
<tr>
<td>Business</td>
<td>$375</td>
<td>Miami</td>
</tr>
</tbody>
</table>
References


