AFHRL-TP-88-58

(2)

AIR FORCE HUMAN RESOURCES LABORATORY

PRINT FORMAT EFFECTS ON ASVAB TEST SCORE PERFORMANCE: LITERATURE REVIEW

Eugene F. Burke
Darrell Hartke
Larry Shadow, Lt Colonel, USAF

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

August 1989

Final Technical Paper for Period October 1986 - November 1988

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

DANIEL L. LEIGHTON, Colonel, USAF
Chief, Manpower and Personnel Division

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AFHRL-TP-88-58	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Manpower and Personnel Division 6b. OFFICE SYMBOL (If applicable) AFHRL/MOAE	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (If applicable) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 62703F; PROJECT NO. 7719; TASK NO. 18; WORK UNIT ACCESSION NO. 46

11. TITLE (Include Security Classification)
Print Format Effects on ASVAB Test Score Performance: Literature Review

12. PERSONAL AUTHOR(S)
Burke, E.F.; Hartke, D.; Shadow, L.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM Oct 86 TO Nov 88	August 1989	20

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	
05	09		aptitude tests; Armed Services Vocational Aptitude Battery (ASVAB); format effects; test construction; test performance
05	08		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Research indicates that differences in the format of answer sheets and test booklets may result in differences in scores on the Armed Services Vocational Aptitude Battery (ASVAB). Because these differences in print format represent a loss in the standardization of test administration, they may be treated as a form of score bias. To address this concern, a literature review was undertaken to identify and integrate research findings and thus determine the utility of future ASVAB research in this area. Few studies directly concerned with test formatting were so identified; however, there is a considerable body of human factors literature which provides a basis for defining variables for future ASVAB format research. In particular, a number of studies supported recent findings of this Laboratory that test scores may be biased by differences in the physical layout of answer sheets and test booklets.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy J. Allin, Chief, STINFO Branch	22b. TELEPHONE (Include Area Code) (512) 536-3877 22c. OFFICE SYMBOL AFHRL/SCV

DD Form 1473, JUN 86 *Previous editions are obsolete.* SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

AFHRL Technical Paper 88-58 August 1989

PRINT FORMAT EFFECTS ON ASVAB TEST SCORE PERFORMANCE: LITERATURE REVIEW

Eugene F. Burke
Darrell Hartke
Larry Shadow, Lt Colonel, USAF

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed by

Linda T. Curran
Acting Chief, Enlisted Selection and Classification Function

Submitted for publication by

Lonnie D. Valentine, Jr.
Chief, Force Acquisition Branch

SUMMARY

A review of the literature of research findings on printing format of test booklets and answer sheets indicates significant impact on test scores. These changes in test scores may have the potential to change or destroy the meaning of test scores for military enlistment use.

PREFACE

This effort was completed as part of our responsibility for improving enlistment selection and classification of Air Force recruits. It was accomplished under work unit number 77191846. The authors wish to express our appreciation to Drs Valentine, Curran, and Ree for guidance during crucial periods of this review and for helpful reviews of this manuscript.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PRINT FORMAT EFFECTS ON ASVAB TEST SCORE PERFORMANCE: LITERATURE REVIEW

I. INTRODUCTION

Bias occurs when systematic errors lead to consistent underestimation or overestimation of statistical or psychometric parameters (Jensen, 1980). A variety of factors may lead to systematic bias in estimating scores on an ability test. Among the most important of these are the conditions under which a test is administered. These conditions include the manner in which items and response alternatives are presented.

In the present report, the term "format" is used to refer to the arrangement of visual information. Formats may vary according to letter/symbol type and size, horizontal and vertical spacing, and line length. These represent the most common variables that have been investigated in research on printed and computer-generated text.

The Armed Services Vocational Aptitude Battery (ASVAB) consists of 10 subtests which measure ability across four principal aptitude domains of quantitative, verbal and clerical abilities, and technical knowledge (Bock & Moore, 1984; Ree, Mullins, Mathews, & Massey, 1982). The individual ASVAB subtests and their respective ability factors are listed below in Table 1.

Table 1. Subtests Comprising the ASVAB

Subtest	Ability Factor
General Science (GS)	Verbal
Paragraph Comprehension (PC)	
Word Knowledge (WK)	
Arithmetic Reasoning (AR)	Quantitative
Mathematics Knowledge (MK)	
Coding Speed (CS)	Clerical Speed
Numerical Operations (NO)	
Auto and Shop Information (AS)	Technical Knowledge
Electronics Information (EI)	
Mechanical Comprehension (MC)	

Recent findings indicate that ASVAB scores are susceptible to differences in both item and answer sheet formats. Sims and Maier (1983) found that male youths in the 1980 National Opinion Research Center (NORC) normative sample achieved significantly lower mean scores than those of male military applicants on the speeded subtests of Numerical Operations (NO) and Coding Speed (CS).

Wegner and Ree (1985) showed that these score differences were due to format differences between the answer sheets used in the NORC 1980 Profile of American Youth study (Bock & Moore, 1984) and those employed in operational ASVAB testing. The answer sheet used in the reference study required examinees to completely fill in circles corresponding to response alternatives, whereas the operational ASVAB answer sheets contained smaller answer blocks requiring only single-stroke entries. Accordingly, the NORC answer sheet was seen to have increased the time needed to respond to the speeded subtest items and, thereby, induced lower test scores.

Welsh and Wegner (1985) found anomalies in mean scores for the ASVAB's Armed Forces Qualification Test (AFQT) composite related to differences in the formatting of NO items. In comparison to ASVAB Forms 11a/b, 12a/b, and 13a/b, the more compact visual style of NO items in ASVAB Form 13c was identified as facilitating higher mean NO scores. The print of Form 13c consisted of a bolder typeface and smaller spacing between items and response alternatives, resulting in faster responses.

To define variables for future study of format bias in ASVAB test scores, a literature review was undertaken. The first stage of this review was a search of the Educational Resources Information Center (ERIC) data base using the keywords "format," "test," "answer sheet," and "text." This search identified a number of human factors studies concerned with the legibility of text and video screen displays, but no test-related citations were found.

A manual search was then undertaken of the psychometrics and test literature. This search focused on the following journals: Applied Psychological Measurement; Educational and Psychological Measurement; Journal of Applied Psychology; Journal of Educational Measurement; and the Psychological Bulletin. Specific issues of other journals were examined as recommended by citations or colleagues. Again, few studies directly concerned with test and answer sheet format were identified.

Based on the results of the original ERIC search, a third search was then undertaken manually. This third and final review of the published literature focused on Ergonomics, Human Factors, and Perceptual and Motor Skills.

The present paper presents a summary and integration of the research findings obtained. In addition to defining potential variables for future ASVAB paper-and-pencil format research, the discussion of format effects is extended to address modal differences introduced by computer-based testing.

II. ANSWER SHEET FORMAT AND TEST PERFORMANCE

Answer sheets offer economy and convenience and generally facilitate test administration and scoring. However, as pointed out by Lindquist (1964), attention should be given to minimizing any adverse effects they may present in terms of test performance and possible distortion of test validity.

The few studies concerned with the effect of answer sheet format on test performance have tended to focus on three independent variables: color, order (vertical versus horizontal grouping of answer spaces), and the shape of response alternative spaces (circles/bubbles versus rectangles/blocks).

The color of the answer sheet does not appear to influence test performance. Michael and Jones (1955) observed no significant differences in college examination scores due to answer sheet color. Similarly, Miller (1965) found no significant effect of answer sheet color on the scores of fourth, eighth, and twelfth grade students on the Verbal Battery of the Lorge-Thorndike Intelligence Test. As will be discussed later in this paper, it is the apparent contrast created by certain color combinations--rather than simply color alone--that results in performance differences.

In a comparison of an International Business Machines (IBM) 805 style answer sheet (horizontally ordered blocks as shown in Figure 1) and eight experimental prototypes, Miller and Minor (1963) required fourth and eighth grade students, college freshmen and sophomores to complete answer sheets in a prescribed sequence (i.e., subjects did not answer test items but followed simple instructions as to which response alternative was to be filled in on the answer sheet). Vertical answer sheet ordering was found to produce fewer correctly marked answer blocks for both the elementary school and college samples. However, Harris

(1986) found no significant effects due to answer sheet ordering for large samples of certification examination scores. This finding was attributed to the robustness of test-taking behavior for adult samples.

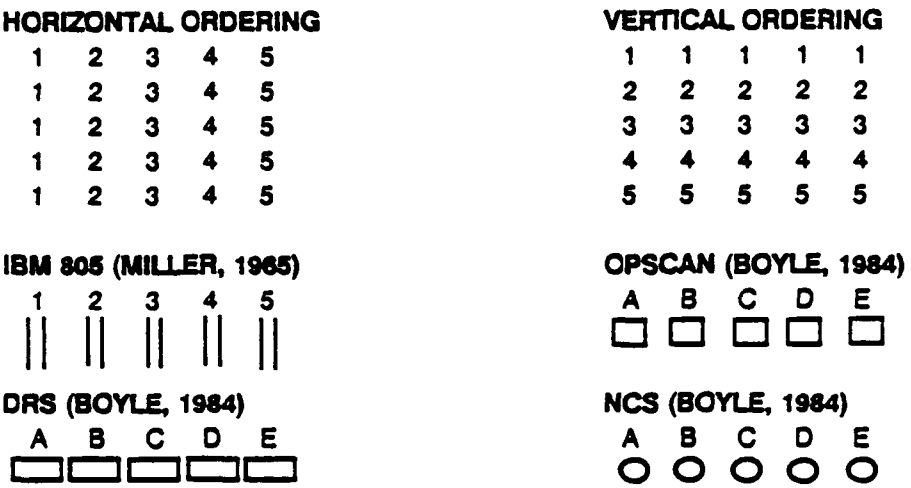

Figure 1. **Ordering and Shape of Answer Sheet Response Blocks.**

Ferguson (1983) found no difference in the level of scores obtained for a much smaller sample of college test scores, but did find that the vertically ordered answer sheet produced a lower standard deviation in scores. Ferguson concluded that the unfamiliar vertical format may have appeared confusing to students and motivated them to reread items and check their responses. Dizney and Davis (1966) compared arithmetic test scores for the IBM 805 and the vertically ordered IBM 1230. Although they found no significant difference in scores, a larger proportion of their sophomore sample reported difficulties with the IBM 1230 (vertical) than with the IBM 805 (horizontal). In a later comparison of the IBM 805 and 1230, Hayward (1967) reported that the expectation of teaching staffs at participating schools was that the horizontally ordered 805 format would be easier to use.

A more definite answer sheet effect appears in relation to the shape of the space for marking responses on the answer sheet. As noted above, Wegner and Ree (1985) concluded that the need to fill in a circular bubble in the NORC study, rather than marking a small rectangular block as in operational ASVAB answer sheets, resulted in degraded performance on the NO and CS speeded subtests. Test equating was therefore reviewed to take this format difference into account in ASVAB norming.

Boyle (1984) noted that the General Aptitude Test Battery (GATB) also employs separate norms according to the type of answer sheet used to record item responses. He compared GATB subtest scores across three answer sheets having either circular bubbles (NCS Form) or rectangles (OPSCAN and DRS Forms) as shown in Figure 1. Significant differences were found for the GATB speeded subtests of Names Comparison and Form Matching, with the bubble format resulting in lower test scores.

Boyle's (1984) study points to the interaction between format and test type; namely, power versus speeded tests, with the latter being subject to answer sheet format effects. Merwin (1963) also found such an interaction in a study involving the Differential Aptitude Test (DAT). He found that only the speeded Coding Speed and Accuracy subtest was subject to answer sheet differences, with a vertical format similar to that of the IBM 1230 yielding lower scores than those achieved with the IBM 805.

Hayward (1967) reported an interaction between answer sheet format and sex. For the unspeeded Sequential Tests of Educational Progress (STEP), three answer sheet formats were compared based on the performance of fourth and eighth grade students whose scores were adjusted for reading level. For the

eighth grade students, females scored higher than males on all three formats, scoring lowest on the IBM 1230. Hayward suggested that females may have better motor coordination and/or coding accuracy than do males, in that the female students encountered no difficulty on any type of answer sheet at fourth grade and scored higher than males at eighth grade.

In their analysis of the 1980 Profile of American Youth data, Bock and Moore (1984) found that NO and CS scores varied significantly by sex but were independent of level of education. Their findings were consistent with other research that indicates female superiority in fine motor movements and repetitive tasks, thereby lending support to Hayward's supposition and pointing to the possibility of interactions among answer sheet format, test type, and sex.

Bock and Moore also found significant differences in ASVAB subtest scores according to level of education achieved which point to the possibility that educational level may also interact with answer format. Perhaps as individuals progress to higher levels of education, they gain more experience in taking tests and greater familiarity with test materials, particularly tests having optically scanned answer sheets such as the Scholastic Aptitude Test (SAT) and the Graduate Record Examination (GRE). Such increased exposure would lead to the more robust test-taking behavior reported by Harris (1986).

To summarize then, performance on paper-and-pencil tests has been found to be influenced by differences in the shape and order of response spaces, with circular bubbles resulting in lower scores and with vertical grouping leading to lower scores and reduced standard deviations. These format variables are most likely to affect speeded tests and may also result in performance differences based on the test-takers' sex and level of education.

III. LEGIBILITY OF PRINTED TEXT

Although much attention has been devoted to writing test items in such a way as to reduce measurement error or bias to minorities, little has been given to the effects of typographical variation in cognitive paper-and-pencil items such as those comprising the ASVAB. In one of the few articles directly concerned with test item format, Vanderplas and Vanderplas (1981) noted that with larger print, there were significant improvements in verbal ability scores for adults ranging in the age from 63 to 85. This study represented a followup to earlier work (Vanderplas & Vanderplas, 1980) in which the researchers had found for older adults improvements in reading speed corresponding to increases in print size.

Although they pertain to only a segment of the general population, these findings point toward a considerable body of human factors research into the format and legibility of printed text. Accordingly, such studies were reviewed to identify format variables likely to act as sources of bias in ASVAB test performance.

A perennial concern with multiversion test batteries such as the ASVAB is the psychometric equivalence of the tests. This section will focus on those parameters which define equivalent legibility of test items as inferred from the human factors literature. Tinker (1963) defined legibility as follows:

> Legibility, then, is concerned with perceiving letters and words, and with the reading of continuous textual material. The shapes of letters must be discriminated, the characteristic word forms perceived, and continuous text read accurately, rapidly, easily, and with understanding . . . . In other words, legibility deals with the coordination of those typographical factors inherent in letters and other symbols, words, and connected textual material which affect ease and speed of reading.

In reading a line of alphanumeric characters, the eye executes a series of jerky ocular movements characterized by three phases: the saccade, the regression, and the fixation (Rayner, 1978; Tinker, 1958).

The saccade serves to bring a new region of text onto the fovea where retinal acuity is highest. Saccades take up approximately 10% of reading time and generally range in length from 2 to 18 characters.

In reading English texts the regression is a right-to-left movement which occurs 10% to 20% of the time in skilled readers. It occurs when the reader has difficulty in understanding text, when text is misinterpreted, or when the reader overshoots his or her target (such as the beginning of a new line).

The fixation phase consists of focusing the eyes on a particular place in the text. Fixations tend to be longer at the start of text, occurring most often five to seven characters from the beginning or end of a line. The amount of material perceived at each fixational pause is referred to as the perceptual span. Optimal typography favors a large perceptual span (i.e., large number of characters/words perceived), whereas nonoptimal text significantly reduces this span. According to Rayner (1978), fixation duration ranges from 100 to over 500 milliseconds (ms). Tinker (1958) indicated the following averages for different types of text: 220 ms for easy prose, 236 ms for scientific prose, 250 ms for adult ordinary reading, and 270 ms to 340 ms for objective test items.

During a fixational pause, the reader must not only see clearly but also comprehend the ideas presented. Accordingly, pause duration includes both perception time and thinking time, and the length of fixation will therefore be affected by the difficulty of the ideas contained in the text. However, reading efficiency at any level of difficulty may also be influenced by typographical arrangement. Nonoptimal text impedes rhythmical perceptual sequences and the perception of words and phrases as whole units. Consequently, such text induces an increase in fixation frequency and duration and an increased number of regressions.

Tinker (1958) cited a study by Lofquist which compared eye movements for a 16-line prose passage and the Minnesota Vocational Test for Clerical Workers (a perceptual speed test requiring comparisons of names and number sequences). The eye movements for the test items reflected a more analytic pattern characterized by more fixations and regressions than those for the prose passage; also, the numerical test items generated the longest perception times. Furthermore, fixation frequency and duration, as well as frequency of regressions, were found to increase with increases in the length, complexity, and difficulty of test items.

Because of the comparatively sophisticated equipment necessary to record eye movements, studies of legibility have tended to use more simple measures as well. Two commonly used measures for assessing legibility are the viewing distance method (whereby the distance between a character and the viewer is adjusted to give that distance at which the character can first be identified unambiguously), and speed reading tests (whereby performance is scored in terms of the number of characters or words read in a fixed amount of time). The results of studies of typographical variation using measures of eye movement, viewing distance, and reading speed are summarized below.

Type Style. The most obvious factor relative to the legibility of printed material is that of the style, face, or font used. Roethlin (1912) reported on the character legibility associated with the width, height, and thickness of line for 10 type faces. Burtt and Basch (1923) found the Cheltenham type face with its uniformly heavy strokes to be more legible than Baskerville and Bodoni (the least legible font in this study). Similarly, Paterson and Tinker (1932) found that American Typewriter and Cloister Black resulted in significant retardation of speed reading scores. Indeed, Tinker (1963) stated that most type faces are about equally legible, and that only extreme deviations in style (such as Cloister Black) significantly affect reading speed. Figure 2 provides examples of type styles from Tinker (1963).

Many factors may play a part, however, in determining the legibility of type. For example, Paterson and Tinker (1932) attributed their results to the inability of typewriter script to adequately adjust the spacing

between characters. Roethlin (1912) attributed hers to differences in apparent contrast across the type faces in her study.

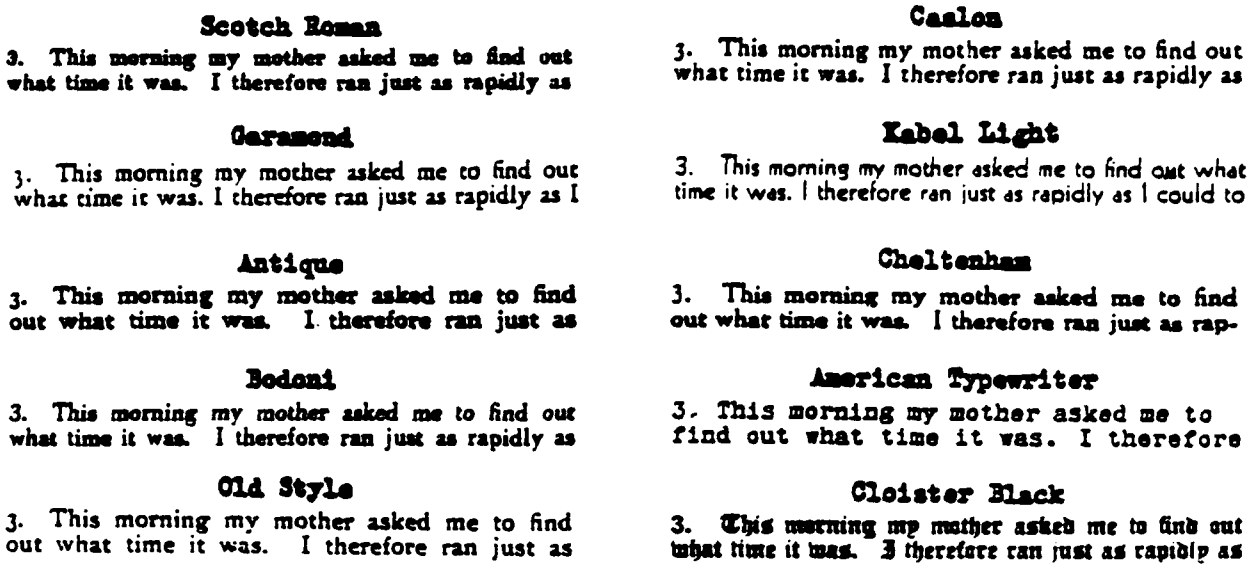

Figure 2. Examples of Type Style (Tinker, 1963).

One of the more consistent results obtained relative to type style is the retarding effect of all-capital text (Poulton, 1967; Poulton & Brown, 1968; Tinker, 1963; Tinker & Paterson, 1955). This finding reflects reading habits accustomed to lowercase text, with all-capital text resulting in an increase in fixation frequency as a result of perception proceeding by individual letter rather than by whole words.

Character Size. The size of the alphanumeric character set places an obvious constraint upon text formatting. Obviously, as pointed out by Smith (1979), a symbol must be large enough to be discriminated from other symbols, but an increase in character size will naturally reduce the total number of characters that can be displayed within a limited area. Various minimum character sizes for printed text were cited by Smith, such as a height of 1.5 millimeters (Fitts, 1951) and 2.3 millimeters for labels (Military Standard 1472B, 1974). However, Smith's accumulation of legibility results suggests a minimum height of 3.5 millimeters at a viewing distance of 0.5 meter.

The size of printed characters is usually defined in terms of the "point," a unit of 1/72 inch used to measure the height of the block of metal on which the letter is cast (examples of different point sizes are given in Figure 3. Tinker (1[illegible] found that legibility and preference increased with letter size between 6-point and 12-point text, with little improvement in legibility beyond 10-point (5/36 inch or 3.5 millimeters). Tinker and Paterson (1955) found that smaller-than-optimal type led to increases in fixation pause duration and fixation frequency as a result of the loss in letter visibility. Larger-than-optimal type (e.g., all-uppercase), since it covers more horizontal space per word, is also characterized by an increase in number of fixations.

Line Length. Variation of line length has a significant effect on fixation frequency. When a line is longer than optimal, it is difficult for the eye to swing back to the beginning of each successive line. The delay that

then occurs interferes with maintaining rhythmical and efficient oculomotor patterns. At the same time, very short lines limit the size of the perceptual span possible and inhibit the use of horizontal peripheral vision.

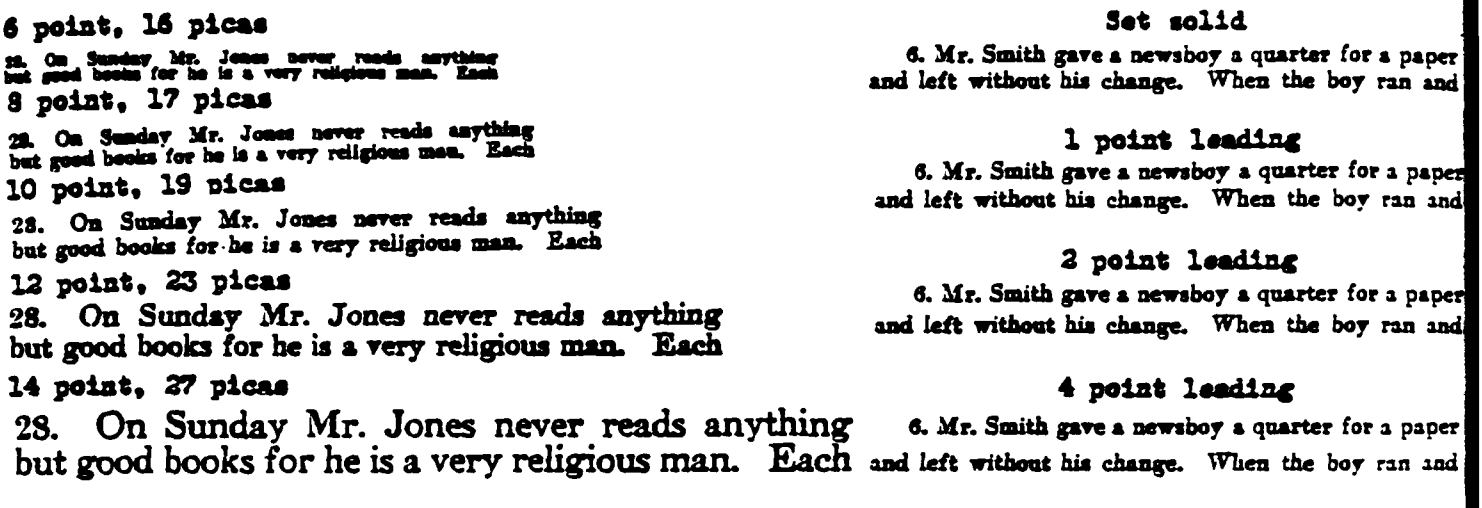

Figure 3. Examples of Type Size, Line Length, and Leading (Tinker, 1963).

In attempting to define optimal line length, Tinker (1963) summarized a number of his own studies in which line length (measured in the pica unit of 1/6 inch or 4.2 millimeters, equivalent to 12-point) was varied along with print size. Although optimal line length varied according to the amount of vertical spacing between lines, Tinker's recommended *minimum* line lengths ranged from 2 1/3 inches (59.3 millimeters) for 6-point type to 2 5/6 inches (72 millimeters) for 12-point type. His maximum line lengths ranged from 4 2/3 inches (118.5 millimeters) to 5 1/2 inches (139.7 millimeters).

Vertical Spacing. As noted above, vertical spacing between lines has a significant effect on legibility. The most extensive investigation of vertical line spacing is that reported by Tinker (1963). His studies varied the spacing between lines in terms of point (1/72 inch) sizes for leading (a printing term for line spacing and representing the small metal strips used to separate blocks of print). When type set solid was compared with 1-, 2-, and 4-point leading, 2-point leading (1/36 inch or 0.7 millimeter) showed an overall superiority, although leading had considerably less influence for 12-point type than for the smaller sizes of 6- and 8-point type. Paterson and Tinker (1932) also found that for 10-point type set with a 19-pica line length, an increase to 4-point leading did not improve ease of reading over 2-point leading.

Miscellaneous. Certainly, there are other factors which may or may not contribute to the optimization of text format. For example, published studies indicate that page margins surrounding printed text do not affect legibility and are therefore redundant. Indenting the first line of a paragraph, however, improves ease of reading. Also, color combinations which increase the brightness contrast between print and paper, as in the use of dark ink on a light color, will enhance legibility. The most legible combinations, according to Tinker, are black-on-white, black-on-yellow, red-on-white, and green-on-red.

Finally, more specific to the ASVAB Numerical Operations (NO) and Coding Speed (CS) subtests, numerals make a greater demand on vision than does normal text comprised of words. As with all-uppercase text, numerals tend to be read digit by digit rather than as whole units, thereby inducing more fixations, longer pause durations, and more regressions. Again, larger sizes are more legible than smaller sizes (8-point versus 6-point), and base numbers used in tables should be printed in boldface (Tinker, 1963).

Tinker's (1963) guidelines for the format of printed text may be summarized as follows:

1. Small type sizes are to be avoided, with an optimal size being 10-point (10/72 inch or 3.5 millimeters) or 11-point type (11/72 inch or 3.9 millimeters).

2. Spacing between lines is particularly important for smaller print sizes, with a 2-point lead (1/36 inch or 0.7 millimeter) being about the most satisfactory.

3. Long and short lines should be avoided. Desirable minimum and maximum lengths vary according to type size and leading, but range between 2 1/3 inches (59.3 millimeters) and 5 1/2 inches (139.7 millimeters).

4. An optimal combination of line length, type size, and leading is 11-point type set in a 22-pica line (3 2/3 inches or 93.1 millimeters) with 2-point leading.

5. Uppercase and lowercase text should be used in preference to all-uppercase. (The latter should be reserved for emphasis, such as in titles.)

6. Color combinations should enhance the brightness contrast between print and page.

IV. LEGIBILITY OF VIDEO TEXT

Recent studies on format have examined factors influencing the legibility and comprehension of computer-generated text. Differences in the mode of test presentation between computer and paper-and-pencil have been found to lead to significant differences in test scores. Given the current development of computer-based versions of ASVAB tests, the results of studies of video-text legibility are reviewed below.

Generally, the results of the earlier studies of format /print appear to generalize to the newer medium of computer-generated text. For example, Pastoor, Schwarz, and Beldie (1983) compared different computer-generated dot matrix character sets based on the Fortune typeface. Smaller character sizes were found to lead to larger performance decrements in tasks requiring a selective strategy (line searching and word identification) than in sequential tasks (reading aloud). In general, higher performance scores and subjective ratings of comfort were obtained for the larger matrix size.

Thus, it would seem that optimal character heights can be defined for video text after the fashion of Tinker's guidelines for print. As for minimum character height, Smith (1979) stated that the most dramatic increase in legibility occurs between visual angles of 0.0015 and 0.003 radian (equivalent to 0.75 and 1.5 millimeters for a viewing distance of 0.5 meter), and that the military standard of 0.0046 radian (2.3 millimeters at 0.5-meter viewing distance) provides a 98% probability of legibility. In their comparison of computer-generated fonts for 5x7 dot matrix characters, Maddox, Burnette, and Gutman (1977) echoed Roethlin's (1912) finding that it is the apparent contrast rather than the font itself which enhances legibility.

The slower reading of video text in comparison to book material has been generally found to be due to the fewer characters per line and the fewer lines per page in the video condition (Kruk & Muter, 1984; Muter, Latremouille, Treurniet, & Beam, 1982), with the slower reading of 40-character lines being due to an increase in fixations (Kolers, Duchnicky, & Ferguson, 1981). Again, speed of reading video text is significantly affected by the vertical spacing of lines. Kruk and Muter (1984) reported that single-spaced text was read 10.9% slower than double-spaced lines, and they recommended that a tight vertical format should be avoided for those computer displays in which the space between lines is small relative to the height of the characters.

The Paterson and Tinker (1932) study cited earlier attributed the poorer legibility of American Typewriter text to the inability of the typewriter to adequately adjust the spacing between characters. Beldie, Pastoor, and Schwarz (1983) found significantly better performance on speed reading and error identification tasks for variable matrix versus fixed matrix character sets. They argued against the use of fixed matrix characters as follows: For wide characters such as M and W, clarity is reduced by fixing a uniform character width; with narrow characters, the relatively large background areas between letters and words create vague word contours.

The effect of variable spacing between words was examined by Trollip and Sales (1986) in their comparison of reading speed and comprehension scores for left-justified (ragged right margin) versus fill-justified text (even left and right margins). Although no significant differences in comprehension scores were observed, they found that subjects took between 11% and 13% longer to read the fill-justified text.

To achieve the even left and right margins of fill-justified text, variations in the length/number of words per line are compensated for by uneven interword spacing which, according to Trollip and Sales, disrupts reading fluency and creates an increase in fixation rate.

Henney (1981) found that though a combination of uppercase and lowercase video text is read faster than all-uppercase, the latter is read more accurately (as would be expected from eye movement studies). Commenting on this, Hathaway (1984) emphasized that reading in schools, newspapers, and magazines sets a normative format using both uppercase and lowercase and that this format should therefore be retained.

Hathaway offered the following guidelines for computer-based instruction:

1. A visual angle of 0.007 radian (a character height of 3.5 millimeters at a viewing distance of 0.5 meter) can be used to guarantee legibility of text.

2. The 80-character line option should be used in preference to 40-character lines.

3. Display text should be double-spaced.

4. Video text should be comprised of both uppercase and lowercase characters.

V. COMPUTER VERSUS PAPER-AND-PENCIL ADMINISTRATION OF THE ASVAB

The previously mentioned AFHRL studies have shown that the speeded ASVAB subtests of Numerical Operations (NO) and Coding Speed (CS) are susceptible to score bias through variations in print format. Format differences in test items themselves affect the ease with which individual items are discriminated from other items in the test booklet. Greaud and Green (1986) found a significant increase in both NO and CS scores for computer versus paper-and-pencil administration which they attribute to the independent presentation of single items and the simplification of the response process by the use of a keypad. For both subtests, subjects found the computer-administered items easier, as indicated by higher scores and shorter response times as compared to the paper-and-pencil versions.

Kiely, Zara, and Weiss (1986) compared paper-and-pencil versions of the NO and CS subtests against two computerized modes of presentation which allowed for both independent presentation of single items and multiple-item presentation scrolled on the screen in blocks. For the NO subtest, the single-item mode was found to yield scores most closely equivalent to paper-and-pencil scores, whereas the multiple-item

mode was found to equate closest to the paper-and-pencil version for CS. Overall, however, the single-item mode of computer presentation yielded higher scores for both subtests.

Evidence also suggests that other ASVAB subtests may be affected by the modality of test presentation. Kiely et al. (1986) compared three scrolling techniques for computer presentation of Paragraph Comprehension (PC) items with standard paper-and-pencil administration. They used a repeated measures design (test-retest) allowing for comparisons across three different sequences of PC presentation: computer followed by paper-and-pencil; paper-and-pencil followed by computer; and paper-and-pencil followed by paper-and-pencil (using parallel versions of the subtest).

The Kiely et al. (1986) results indicate the problems that can be expected in attempting to emulate the relatively complex format of a test that has high text content. Significant differences were found between computer versus paper-and-pencil administration of PC (with the computer-based version yielding lower test scores); for type of scrolling technique (with the entire screen for viewing text technique yielding the lowest scores); and for sequence of presentation (with the computer version prior to paper-and-pencil administration yielding the lower scores). The authors suggested that the effect observed for scrolling mode resulted from differences in the information and memory load created by particular screen configurations. They also suggested that the asymmetric transfer effect observed among administrative sequences was due to a possible increase in test anxiety induced by a combination of the complexity of the PC subtest and lack of familiarity with computers.

Although Kiely et al. (1986) found no substantial differences between paper-and-pencil and computer-based versions of the power subtests of Auto and Shop Information (AS), Electrical Information (EI), and Mechanical Knowledge (MK), Lee, Moreno, and Sympson (1986) found a small but significantly lower mean score for a computer adaptive test (CAT) version of the Arithmetic Reasoning (AR) subtest compared to the paper-and-pencil version. In contrast to the results of Greaud and Green (1986), who found improved scores for the computerized versions of NO and CS, Lee et al. found that 21 of the 30 computer-administered AR items had lower scores compared to the paper-and-pencil version.

Thus, significant format effects on ASVAB subtest scores have been found to occur with changes in the mode of presentation from paper-and-pencil to computer.

VI. DISCUSSION

Despite the lack of research concentrated upon the effects of format variation on cognitive test performance, it is readily apparent from the preceding review that certain format variables offer potential for future ASVAB research.

These variables are as follows: answer block shape (answer block ordering might also be included to determine whether ordering accentuates the effect of shape); character size (in points/millimeters); and vertical spacing (or leading, which has been shown to have a greater influence on legibility when smaller type sizes such as 6- or 8-point are used). Where typographical effects on prose items are examined, line length variations should also be investigated.

Selection of a subset of the 10 ASVAB subtests will be necessary to reduce the format effect research to manageable and practical proportions. The diagrammatical content of such subtests as AS and MK places them beyond the focus of the present review, which is upon prose and numerical text. Furthermore, research has indicated that these tests may be relatively robust with respect to format variations (Kiely et al., 1986). Given the results of prior research and as speeded components of the ASVAB, NO and CS obviously warrant inclusion. With regard to the power components of the ASVAB, the PC and AR subtests provide potentially

suitable media for analysis of typographical variation of items by virtue of their prose structure and the results summarized in the previous section.

Accordingly, scores on AR, CS, NO, and PC could be treated as dependent variables in observing the effects of format variation. Given the differences in the format characteristics of these four subtests (e.g., PC versus CS), the experimental designs for studying format effects will vary according to test type. Whereas a crossed design would be suitable for manipulating character size and vertical spacing of CS items, inclusion of line length in the variation of PC items would suggest a nested design. Manipulation of answer sheet format could follow the split-plot factorial design employed by Boyle (1984).

Kirk (1968) offered the prescription that it is generally more efficient to proceed with a "... sequence of relatively small experiments, each based on the results of the preceding experiment" (p. 244). Experiments concerned with specific subtest and answer sheet formats could therefore be conducted according to a hierarchically ordered sequence in which only those format variables found to have significant effects are included in subsequent combinations across subtests and answer sheets. Furthermore, the results of this sequence of mini-studies could then be combined with more traditional bias variables to identify interactions between format variation and education, ethnicity, and gender; this might be particularly beneficial if done in conjunction with the currently ongoing studies of the effect of the latter variables on ASVAB subtest scores.

Because the effects of variation in typographical layout have been found to generalize to computer-generated text, examination of the effects of these variables on paper-and-pencil test performance would be of considerable utility to current efforts to automate the ASVAB. Where typographical variables are found to affect only paper-and-pencil or computerized administration and not both, a clearer distinction would be achieved in determining model differences in test presentation.

Precise specification of experimental designs for studying format variation requires the consideration of statistical parameters of effect size (e.g., Type I and Type II error rates) so that optimum sample sizes can be defined (Cohen, 1977).

REFERENCES

Beidle, I.P., Pastoor, S., & Schwarz, E. (1983). Fixed versus variable letter width for televised text. *Human Factors, 25*, 273-277.

Bock, R.D., & Moore, E.G.J. (1984). *Profile of American Youth: Demographic influences on ASVAB test performance*. Chicago: National Opinion Research Center.

Boyle, S. (1984). The effect of variations in answer sheet format on aptitude test performance. *Journal of Occupational Psychology, 57*, 323-326.

Burtt, H.E., & Basch, C. (1923). Legibility of Bodoni, Baskerville Roman, and Cheltenham type faces. *Journal of Applied Psychology, 7*, 237-245.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Dizney, H.F., & Davis, O.L. (1966). Effects of answer sheet format on arithmetic test scores. *Educational and Psychological Measurement, 26*, 491-493.

Ferguson, W.F. (1983). *Non-traditional answer sheet format: Solution or problem*? Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Nashville, TN.

Fitts, P.M. (1951). Engineering psychology and equipment design. In Stevens, S.S. (Ed.), *Handbook of Experimental Psychology*. New York: Wiley.

Greaud, V.A., & Green, B.F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10*, 23-34.

Harris, D.J. (1986). A comparison of two answer sheet formats. *Educational and Psychological Measurement*, 46, 475-478.

Hathaway, M.D. (1984). Variables of computer screen display and how they affect learning. *Educational Technology, 1*, 7-11.

Hayward, P.C. (1967). A comparison of test performance on three answer sheet formats. *Educational and Psychological Measurement, 27*, 991-1104.

Henney, M. (1981). *The effects of all-capital print versus regular mixed print, as displayed on a microcomputer screen, on reading speed and accuracy* (ERIC Document No., ED 208 359).

Jensen, A.R. (1980). *Bias in mental testing*. New York: The Free Press.

Kiely, G.L., Zara, A.R., & Weiss, D.J. (1986). *Equivalence of computer and paper-and-pencil Armed Services Vocational Aptitude Battery tests* (AFHRL-TP-86-13, AD-A171 187). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Kirk, R.E. (1968). *Experimental design*. Monterey, CA: Brooks/Cole.

Kolers, P.A., Duchnicky, Y., & Ferguson, D.C. (1981). Eye movement measurement of readability of CRT displays. *Human Factors, 23*, 517-527.

Kruk, R.S., & Muter, P. (1984). Reading of continuous text on video screen. *Human Factors, 26*, 339-345.

Lee, J.A., Moreno, K.E., & Sympson, J.B. (1986). The effects of mode of test administration on test performance. *Educational and Psychological Measurement, 46*, 467-473.

Lindquist, E.F. (1964). *Basic considerations in answer sheet design*. Houghton Mifflin Company Circular.

Maddox, M.E., Burnette, J.T., & Gutmann, J.G. (1977). Font comparisons for 5 x 7 dot matrix characters. *Human Factors, 19*, 89-93.

Merwin, J.C. (1963). New measurement research center answer sheets and Differential Aptitude Test norms. *Student Counseling Bureau Newsletter*. Minneapolis: Office of Dean of Students, University of Minneapolis, April 15th.

Michael, W.B., & Jones, R.A. (1955). The influence of color of paper upon scores earned on objective achievement examination. *Journal of Applied Psychology, 39*, 447-450.

MIL-STD-1472B. (1974, December). *Human engineering design criteria for military systems, equipment and facilities*. Washington, DC: U.S. Department of Defense.

Miller, I. (1965). A note on the evaluation of a new answer form. *Journal of Applied Psychology, 49*, 199-201.

Miller, I., & Miner, F.J. (1963). Influence of multiple-choice answer form design on answer-marking performance. *Journal of Applied Psychology, 47*, 374-379.

Muter, P., Latremouille, S.A., Treurniet, W.C., & Beam, P. (1982). Extended reading of continuous text on television screens. *Human Factors, 24*, 501-508.

Pastoor, S., Schwarz, E., & Beldie, I.P. (1983). The relative suitability of four dot-matrix sizes for text presentation on color television screens. *Human Factors, 25*, 265-272.

Paterson, D.G., & Tinker, M.A. (1932). Studies of typographical factors influencing speed of reading. *Journal of Applied Psychology, 16*, 605-613.

Poulton, E.C. (1967). Searching for newspaper headlines printed in capitals or lower-case letters. *Journal of Applied Psychology, 51*, 417-425.

Poulton. E.C. (1972). Size, style and vertical spacing in the legibility of small typefaces. *Journal of Applied Psychology, 56*, 156-161.

Poulton, E.C., & Brown, C.H. (1968). Rate of comprehension of an existing teleprinter output and of possible alternatives. *Journal of Applied Psychology, 52*, 16-21.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85*, 618-660.

Ree, M.J., Mullins, C.J., Mathews, J.J., & Massey, R.H. (1982). ASVAB: *Item and factor analyses of Forms 8, 9, and 10* (AFHRL-TR- 81-55, AD-A113 465). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Roethlin, B.E. (1912). The relative legibility of different faces of printing styles. *American Journal of Psychology*, 23, 1-36.

Sims, W.H., & Maier, M.H. (1983). *The appropriateness for military applications of the ASVAB subtests and score scale in the new 1980 reference population.* Alexandria, VA: Center for Naval Analyses.

Smith, S.L. (1979). Letter size and legibility. *Human Factors, 21*, 661-670.

Tinker, M.A. (1958). Recent studies of eye movements in reading. *Psychological Bulletin, 55*, 215-231.

Tinker, M.A. (1963). *Legibility of print.* Ames, IA: Iowa State University Press.

Tinker, M.A., & Paterson, D.G. (1955). *The effect of typographical variations upon eye movement in reading. Journal of Educational Research, 49*, 171-184.

Trollip, S.R., & Sales, G. (1986). Readability of computer-generated fill-justified text. *Human Factors, 28*, 159-163.

Vanderplas, J.M., & Vanderplas, J.H. (1980). Some factors affecting legibility of printed materials for older adults. *Perceptual and Motor Skills, 50*, 923-932.

Vanderplas, J.M., & Vanderplas, J.H. (1981). Effects of legibility on verbal test performance of older adults. *Perceptual and Motor Skills, 53*, 183-186.

Wegner, T.G., & Ree, M.J. (1985). *Armed Services Vocational Aptitude Battery: Correcting the speeded subtests for the 1980 youth population* (AFHRL-TR-85-14, AD-A158 823). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Welsh, J.R., & Wegner, T.G. (1985). *Initial operational test and evaluation of Armed Services Vocational Aptitude Battery (ASVAB) Forms 11, 12, and 13: Data quality analysis.* Paper presented at the 27th Annual Conference of the Military Testing Association, San Diego, CA.