DOCUMENT RESUME

ED 059 251 TM 001 065

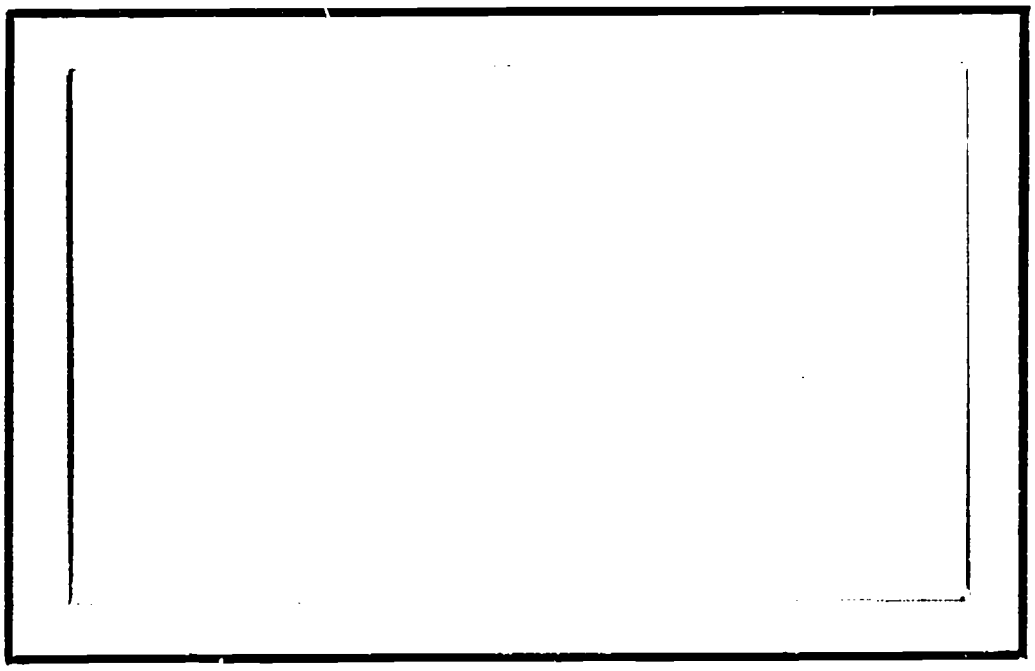
AUTHOR Veldman, Donald J.; And Others
TITLE Computer Scoring of Sentence Completion Data.
INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
REPORT NO RMM-3
PUB DATE 68
CONTRACT OEC-6-10-108
NOTE 21p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Attitudes; Computer Oriented Programs; *Computer Programs; Data Analysis; Data Processing; *Evaluation Techniques; Participant Characteristics; *Personality Assessment; Personality Studies; Personality Tests; Research Methodology; *Scoring; Scoring Formulas; Semantics; Sentences; Sex Differences; Statistical Analysis; *Verbal Tests
IDENTIFIERS *One-Word Sentence Completion

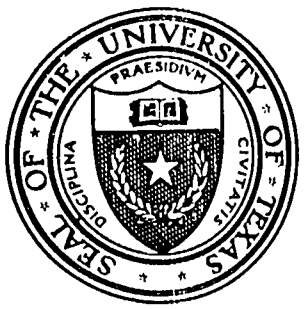
ABSTRACT

This paper outlines the development of techniques for computer-based personality assessment from sentence completions. The One-Word Sentence Completion (OWSC) instrument was designed to elicit data suitable for machine processing, while retaining most of the advantages of a free-response format. Two operative scoring systems are described. The first employs a "dictionary" of 4366 weighted response words to yield 25 scores from a 90-item OWSC form. The second system utilizes a complex word-root data reduction procedure and a bank of 892 generic roots to produce scores for 40 variables. Initial reliability data and normative sex differences are reported, and future development of the technique is discussed. (Author/DLG)

N-AI



The Research & Development Center For Teacher Education



THE UNIVERSITY OF TEXAS

AUSTIN

RMM-3

COMPUTER SCORING OF SENTENCE COMPLETION DATA

Donald J. Veldman
Shirley L. Menaker
Robert F. Peck

Spring, 1968

COMPUTER SCORING OF SENTENCE COMPLETION DATA

Donald J. Veldman, Shirley L. Menaker

and Robert F. Peck

The University of Texas at Austin

The development of techniques for computer-based personality assessment from sentence completions is outlined. The One-Word Sentence Completion (OWSC) instrument was designed to elicit data suitable for machine processing, while retaining most of the advantages of a free-response format. Two operative scoring systems are described. The first employs a "dictionary" of 4366 weighted response words to yield 25 scores from a 90-item OWSC form. The second system utilizes a complex word-root data reduction procedure and a bank of 892 generic roots to produce scores for 40 variables. Initial reliability data and normative sex differences are reported, and future development of the technique is discussed.

Published in Behavioral Science, 1969, 14, 501-507.

The development of this technique was supported in part by the Computer Analysis of Personality project, National Institute of Mental Health Grant No. 06823, and by the Research and Development Center for Teacher Education, U.S. Office of Education contract OE-6-10-108.

COMPUTER SCORING OF SENTENCE COMPLETION DATA

Donald J. Veldman, Shirley L. Menaker

and Robert F. Peck

The University of Texas at Austin

Although computer programs have been written to generate verbal summaries of the numerical-score profiles produced by a variety of personality questionnaires such as the MMPI (Swenson, et al., 1965), few attempts have been made to accomplish the obverse task -- generating quantitative indices from verbal responses to ambiguous stimuli (Veldman & Menaker, 1968). At the present time, the authors are aware of only three major research programs which are concerned with computer applications of this type: Stone's (1966) General Inquirer system for content analysis of verbal text, Gorham's (1967) procedures for scoring responses to the Holtzman Inkblot Technique, and our own work with the One-Word Sentence Completion method, which will be described in this article.

The General Inquirer is virtually unique as a research tool for behavioral scientists. By the use of a pre-categorized dictionary of words and idiomatic phrases, and this system of computer programs, it is possible to derive quantitative indices for an endless variety of psychological constructs by completely objective processing of any verbal text. The system has been used most often with narrative materials, but experimental applications have been made to other kinds of data, such as sentence completions (Goldberg, 1967).

Another very successful effort to accomplish computer scoring of verbal data was the development by Mosely (1963) and Gorham (1967) of a procedure for handling six-word responses to each of the 45 stimuli of the Holtzman Inkblot Technique. Remarkably close agreement has been obtained between computer-based

scoring of protocols from group administration, and hand scoring by trained examiners in individual testing of the same subjects.

In common with the purposes of Stone and Gorham, we have sought to use computer programs for interpretation of sentence-completion responses for three major reasons. First of all, the computer time needed for scoring a protocol is far less expensive than that of an experienced clinical psychologist. An effective computer scoring program would make use of verbal free-response instruments much more practical in large-scale research operations. The second reason is the complete objectivity of computer procedures. Unlike human judges, whose interpretive behavior is capricious at worst and idiosyncratic at best, a computer will follow the rules embodied in its program with utter faithfulness. The third and most interesting reason for trying to program a computer to score sentence completions is the heuristic value of the programming process itself. Vague, incomplete, or internally contradictory interpretive procedures simply cannot be programmed. Unlike a human judge, a computer cannot tolerate ambiguity in the rules upon which it operates. As experienced clinicians with typical self-assurance about the logical basis of our interpretive behavior, we have been rather chagrined at times by our own inability to state operationally just how we arrive at particular conclusions on the basis of certain responses. Because a computer program operationally defines an interpretive theory, it will be possible to test many of the "rules" that are now only clinical lore.

The One-Word Sentence Completion Method

When we began working in this area in 1961, we were very much aware of the limitations of computer storage devices, and the difficulties of interpretation posed by the vagaries of English syntax. To avoid these problems and

still retain the open-end quality of the sentence completion format, we designed a 90-item One-Word Sentence Completion (OWSC) form with instructions to complete each of the sentence stems by inserting a single word. Since then, similar forms containing 50 and 35 stems have been designed, and protocols have been gathered from roughly 10,000 high school and college students in Texas. About 800 Spanish-language protocols have also been obtained in Mexico, Venezuela, and Chile.

The incomplete sentences which comprise these forms were purposely designed to sample a wide variety of topics and formats, since our purpose was to explore the possibilities of the technique. Some stems require descriptor responses ("My father is _____."); others present reactions and ask for stimulus objects ("_____ makes me angry."); and still others draw transitive verbs ("Men often _____ women.")

Although these protocols were intended for machine processing, they have been used in some situations as part of a battery for individual clinical assessment, and seem to retain most of the value of the usual free-response sentence completion format. Although one would expect to lose the information contained in the extent of the subjects' responses, much of it seems to remain in the frequency of blanks and bland response words under the single word restriction.

Data Preparation Procedures

Because we realized at the outset that some aspects of the data would have to be sacrificed to permit efficient reduction prior to the actual scoring process, we established the following conventions for transferring the raw responses to punch cards.

(1) All misspellings were corrected. We have counted them for some studies, however, and added this information to the records.

(2) Hyphens were removed and, when multi-word responses were given despite the instructions, the spaces were ignored and the response characters were punched continuously.

(3) Private proper-name responses ("John makes me happy.") were all coded "PN," but public proper names were punched verbatim.

(4) Responses were packed up to a 16-character limit. On the basis of our first attempts at scoring, this limit was later reduced to ten characters. Machine characteristics had some influence here, also: The 48-bit word of the CDC 1604 allowed us to store a 16-character response in two memory locations; at present we can store a 10-character response in one location of the CDC memory.

After key-punching, the raw data were transferred to magnetic tape, and lists of all different responses were compiled separately for each stem. All programming was done in FORTRAN, and most of the basic procedures are described in a recent book by Veldman (1967). The compiled lists became the focus for most of our initial attempts to develop scoring procedures. More recently we have developed methods for reducing these response banks to generic forms, fewer in number but more general in applicability.

Early Attempts to Design Scoring Systems

The concept that led us into this research was that of language translation. We began with the vague notion that sentence completion data -- in the language of the subject -- could be translated by machine into personality descriptions -- in the language of the psychologist. After our first efforts

to operationalize this idea resulted in exasperated confusion, we enlisted the aid of professional logicians in hopes of "mapping the interpretive space." Although some progress was made in clarifying our thinking about the interpretive process, it eventually became quite obvious that the task of outlining in sufficient detail the incredibly complex tree-structure of a clinician's potential behavior when faced with a protocol was far beyond the limits of computer memories or our collective patience.

Perhaps in reaction to this impasse, we next devoted our efforts to devising an empirical prediction system involving no theoretical basis at all. Criterion groups were selected and the computer was programmed to determine optimal weights for every different response to each stem. Cross-validation studies indicated moderate success with this approach, with certain criteria. This "black box" method, however, was inherently limited in that it yielded no information about how it achieved its successes or why it failed when it did. Some of the data collected during this phase of the project strongly suggested that further reduction of the raw data to higher-order categories might substantially improve the efficiency of the system.

Clinical Response Weighting

At the same time that the empirical scoring approach was being explored, Shirley Menaker pursued an entirely different line of attack. After studying the lists of raw responses to each stem which had been compiled from a sample of 1000 female sophomores enrolled in the College of Education, she selected 25 psychological variables for which sufficient information appeared to be provided by the 90-item OWSC form.

The stems were divided into two classes with regard to the 25 variables: Primary stems for a given variable were those for which all responses could be weighted. For example, stem 36, "My mother _____ me," was a primary stem for variable 9, Perception of Mother. Secondary stems for a particular variable were those which occasionally elicited relevant data. For instance, the response "mother" to stem 41, "When I need help, I usually depend on _____," was given a positive weight for variable 9, but other responses were ignored. Primary stems were available in the 90-item OWSC form for all 25 constructs.

The original raw response total was, of course, 90,000. By compilation of stem lists, the total was reduced to 16,829. When idiosyncratic responses were ignored, the codable data was reduced to 7,142 responses. Dr. Menaker spent approximately 100 hours assigning weights to 4,366 of these non-unique responses, using a 17-point continuum for each variable. Of the 7,142 words, 2776 were considered neutral with regard to all variables and were not included in the system.

The 4,366 alphabetic responses with the stem numbers, variable numbers, and assigned weights were punched and a program was written to score individual protocols, which an IBM 7040 computer processed as fast as it could read the data cards.

To evaluate the comparability of the machine scores and clinical ratings on the same 25 variables, two judges experienced in the use of sentence completion data were asked to rate the original OWSC protocols of a sample of 79 female elementary education majors who had not been included in the basic sample of 1000 cases. The judges used 7-point rating scales and a brief manual

which gave a few examples of high, middle, and low responses for each variable. Table 1 contains a list of the 25 variables with the correlation coefficients that were obtained between the two judges, and between the computer scores and the sum of the two judges' ratings. It is apparent that the computer agreed with the judges as well as they agreed with each other.

Table 1

Inter-judge and Computer-Judge Correlations for 25 Variables Comprising the Initial Scoring System (N=79)

Variable	Between Judges	Computer vs. Judges
1. General Self Perception	.51	.46
2. Optimism - Pessimism	.61	.75
3. Sexual Self Perception	.75	.76
4. Psychosexual Integration	.48	.71
5. Attitude toward Own Past	.76	.78
6. Independence, Self-Reliance	.62	.69
7. Confidence re Classroom Discipline	.63	.47
8. Attitude toward Father	.49	.53
9. Attitude toward Mother	.70	.56
10. Attitude toward Men	.72	.73
11. Attitude toward Women	.63	.59
12. General Attitude toward Others	.72	.60
13. Extraversion - Introversion	.38	.45
14. Attitude toward Authority	.65	.63

15.	Implied Teacher-Child Interaction	.59	.51
16.	Self in Parental Role	.91	.94
17.	Attitude toward Teaching Profession	.93	.80
18.	Self in Marriage Role	.95	.89
19.	Attitude toward Stress	.87	.76
20.	Persistence, Tenacity	.56	.55
21.	Perception of Own Ability	.50	.57
22.	Intellectual Concern	.50	.53
23.	Clarity re Future	.67	.76
24.	Energy, Activity Level	.69	.82
25.	General Mental Health	.70	.68

Extensive validity studies of this scoring system are now in process. Preliminary evidence indicates that the computer-derived scores are as useful as ratings made by clinicians -- and far less expensive.

Data Reduction to Generic Roots

Although effective and simple enough to be used with relatively small computers, the scoring system described above has certain weaknesses. The most serious of these is the relatively small proportion of the raw data which is actually utilized. All unique responses, for instance, are excluded from consideration, and over 10% of the responses in the sample of 1000 protocols were idiosyncratic. Also, for any variable with an actual mean other than zero, such idiosyncratic responses are implicitly mis-weighted.

In order to include a larger proportion of the idiosyncratic responses, and at the same time increase the generality of the system to new subject samples,

a data-reduction system was designed to go beyond the compilation of lists of different responses for each stem. The data used were obtained from 2321 freshmen (1362 males and 959 females) who completed a 36-item OWSC form as part of an institutional research project at the University of Texas.

Stage One. Keypunching and Compilation of Raw Responses. The conventions described earlier were followed, and responses were punched to a limit of 10 characters. Lists of all non-unique responses were compiled. Of the 83,556 responses, 1.6% were blanks and 9.5% were unique. By compilation of identical responses, the original data were reduced 93% to a list of 5772 non-unique responses.

Stage Two. Reduction to Word Roots. The 5772 responses were listed alphabetically and inconsequential suffixes as well as common prefixes were eliminated. This left a list of 1700 root forms. For example, the root form LOV was retained and LOVE, LOVES, LOVED, LOVING, etc., were eliminated. The term LOVELY was retained, however, because of its different semantic implication. A FORTRAN routine was then constructed to use this root list to carry out the reduction process on any raw response input to it. Figure 1 describes the procedure. Application of this routine to the raw data reduced the number of uniques from 9.5% to less than 2% of the raw data.

Stage Three. Grouping to Define Generic Roots. By clustering roots which were clearly synonymous, the list of 1700 roots was further reduced to a total of 892 "generic" roots. For each of these, one other of the 892 was designated as its semantic opposite. A higher-order routine was then written to (1) input the raw response, (2) find its word root, and (3) output the appropriate direct or negation generic form code.

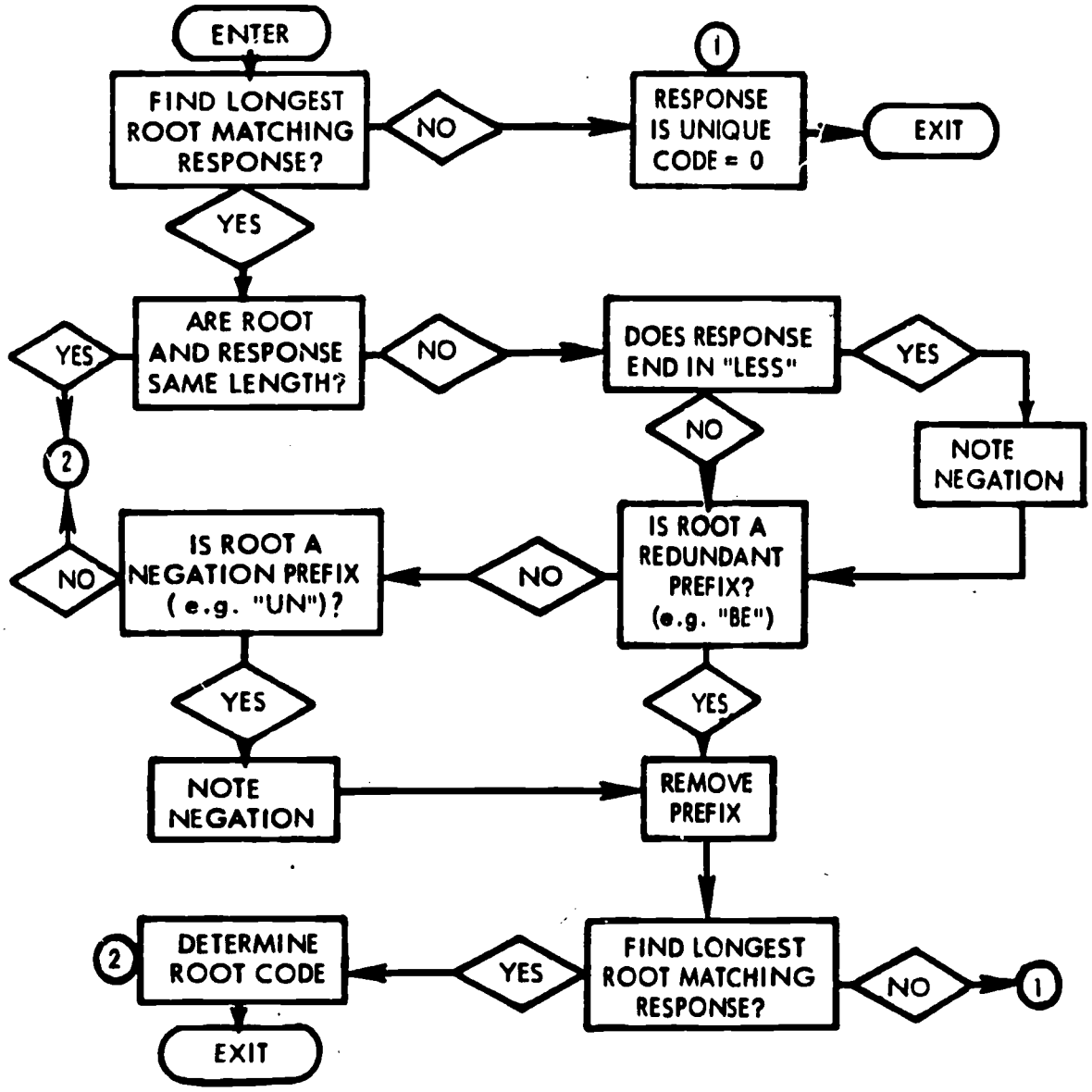


FIGURE 1. REDUCTION OF RESPONSES TO WORD-ROOT FORMS.

Stage Four. Construction of a Generic Root Scoring System. The compiled lists of raw responses to each stem, which were stored on tape in Stage One of the processing, were converted to generic root forms and were then recompiled and printed. The five-digit codes (1-2=stem, 3-5=generic word root) representing the various stem-generic combinations appearing in these lists were then used to define a series of 40 structural and psychological variables. Thirteen of these variables were simple counts of the occurrence of particular types of generics, or of other protocol characteristics (see Table 2), while the other 27 variables were defined by two sets of stem-generic combination codes. For instance, the variable called Optimism (23) was scored by counting the frequency of occurence of a group of stem-generic codes representing negative expectations for the future, and subtracting this value from a frequency count of occurrences of another group of stem-generic combinations representing positive expectations. Almost every stem-generic combination which appeared in the recompiled lists was assigned to one or another scoring category. The few that were not used were either ambiguous or considered neutral with regard to all variables.

Table 2

Male-Female Differences on the Variables of the Word-Root Scoring System[a]

Variable	Male $\bar{X}$	Female $\bar{X}$	P Level
1. Response Length. Average number of characters per response.	6.66	6.74	.003
2. Response Variation. Number of different generics used in protocol.	28.97	29.50	.0003
3. Consistency Index. Identical responses to four pairs of stems.	1.36	1.22	.0005
4. Pronoun Responses. Non-self referent. (to all stems)	.09	.06	.01
5. Evasive Responses. Stem repetitions and other deliberate evasions.	1.62	1.26	<.0001
6. Proper Names. Pre-coded "PN". (to all stems)	.12	.12	NS
7. Non-Responses. Blanks. (to all stems)	.64	.51	NS
8. Self References. Pronouns. (to all stems)	.19	.21	NS
9. Age Responses. (to all stems)	.25	.22	NS
10. Orality. References to food, eating, drinking, or smoking.	.12	.07	.004
11. Money. References to acquisition, use, or lack of money.	.37	.17	<.0001
12. Ideology. References to politics or religion.	.31	.32	NS
13. Commonality. Ordinary, simple vs. unusual, complex.	.18	.13	NS
14. Somatic Self Esteem. Strength, health, attractiveness.	.30	.23	.04
15. Social Self Esteem. Extraversion, self-confidence.	.51	.49	NS

Variable	Male $\bar{X}$	Female $\bar{X}$	P Level
16. Cognitive Self Esteem. Ability, intelligence.	.43	.17	<.0001
17. Performance Self Esteem. Success or improvement.	-.60	-.48	.004
18. General Self Esteem. Primarily the sum of variables 14-17.	.65	.39	.01
19. Emotional Stability. Control, mildness of emotions.	-.03	-.26	<.0001
20. Character. Morality, responsibility, behavioral control.	.42	.68	<.0001
21. Impulse Acceptance. ("When an animal is wild, it is _____.")	.12	-.02	.0003
22. General Mood. Happiness and satisfaction.	.29	.19	.05
23. Optimism. Expectations for the future.	.46	.46	NS
24. Certainty. Clarity and decisiveness in general.	.03	-.06	NS
25. Self-Confidence. Calm, brave vs. anxious, fearful.	-.05	-.22	.0001
26. Ambiguity Acceptance. ("Darkness is _____.")	.15	.09	.04
27. Stress Resistance. ("I _____ when put under pressure.")	.15	-.14	<.0001
28. Academic Attitude. Cathexis of school, studying.	.25	.26	NS
29. General Motivation. Ambition, effort, interest.	2.25	2.01	.002

Variable	Male $\bar{X}$	Female $\bar{X}$	P Level
30. Attitude toward Mother. ("My mother is _____.")	.67	.67	NS
31. Attitude toward Father. ("My father is _____.")	.62	.62	NS
32. Attitude toward Family. Primarily the sum of variables 30 and 31.	1.28	1.29	NS
33. Attitude toward Men. ("Most men are _____.")	.14	.25	.001
34. Men toward Women. ("Men often _____ women.")	.36	.20	<.0001
35. Attitude toward Women. ("Most women are _____.")	.29	-.04	<.0001
36. Women toward Men. ("Women often _____ men.")	.11	.17	NS
37. Heterosexuality. Cathexis of opposite sex and marriage.	1.12	1.55	<.0001
38. Attitude toward Average Person. ("The average person is _____.")	.26	.31	NS
39. General Interpersonality. Friendly. kind, courteous.	1.37	1.82	<.0001
40. Unique Responses. Responses for which root forms could not be found.	.77	.53	<.0001

[a]N=1362 males, 959 females

Variables 13-39 had bipolar definitions.

Table 2 lists the 40 variables for which the current scoring program yields quantitative values. As implied by the descriptions in this table, some of the variables are based on data from only one stem, although most of the variables utilize information from a variety of the 36 items on the OWSC form.

Table 2 also contains the means and the significant (≤5%) p-levels for tests of the differences between the male and female sub-samples. In summary, the females appear to be more verbally fluent, while the males employ more evasive and non-commital responses. The males report greater self-esteem, self-confidence, emotional stability, and resistance to stress, while the females indicate somewhat more positive attitudes toward other people and are more positively oriented toward marriage. The males indicate more positive general motivation, but not in the academic area considered alone. The males are more concerned about money, and obtain higher orality scores than do the females. The males indicate higher cognitive self-esteem (ability), but lower performance self-esteem. Although direct evaluations are higher toward the opposite sex, indirect sex-allegiance appears in expectations for heterosexual relationships.

Construct validation of these variables is now in process, utilizing a variety of self-report attitude data and academic performance measures that are available for various student samples. Further refinement of the scoring system through re-definition of variables in terms of the stem-generic combinations that are assigned to them is also planned as validity data are acquired.

This system requires a rather large computer to handle the storage of the word-roots and the stem-generic code lists for the 40 variables (about 15,000 DIMENSIONED locations). Further experience with the technique may allow

a substantial reduction in the current lists of 1700 word roots and 892 generic roots, thus reducing memory requirements and processing time.

Directions of Future Research

Beyond the refinement of the score definitions based on validation evidence, we intend to explore further the empirical determination of optimum utilization of the data for particular diagnostic purposes. Using generic word roots rather than raw responses in this manner may reveal useful characteristics of the verbal behavior of individuals which our present system ignores.

The design of a new OWSC form which includes stems that will systematically sample from a theoretically determined "assessment space" is another of our goals for the near future. Although the present forms cover a wide variety of attitude and personality concepts, they do not do so on the basis of any a-priori scheme. The definition of scoring variables should be simplified and improved by the use of a form designed in this way.

Finally, we hope to continue the development of computer-based assessment systems that interact on-line with a psychologist or with the subject himself. An exploratory study of the latter (Veldman, 1967b) indicated that a sentance completion procedure with a computer-controlled "inquiry" can in many cases clear up the ambiguity of a subject's responses, and even yield a "second level" of data under some conditions.

The other aspect of the potential for man-machine interaction using time-shared remote consoles is the possibility of a two-stage cooperative assessment procedure. The computer would be fed the raw responses of the subject and would proceed with an analysis of their implications using a large normative data base. When it encountered idiosyncratic words, it would ask the psychologist

for synonyms. It would produce a sort of "laboratory report" of its findings with regard to major personality dimensions, and would also call to the psychologist's attention particular features of the protocol which, being normatively rare, might have special interpretive significance beyond the scope of the machine's general data base. By making the most of the machine's ability to systematically extrapolate from large-scale normative "experience," and the human clinician's unique ability to interpret by analogy, the quality of personality assessment could be greatly improved over our present reliance on machine-scored questionnaires and clinically-interpreted projective protocols.

References

Goldberg, J. B. Computer analysis of sentence completions. Journal of Projective Techniques and Personality Assessment, 1966, 30(1) 37-45.

Gorham, D. R. Validity and reliability studies of a computer-based scoring system for inkblot responses. Journal of Consulting Psychology, 1967, 31(1), 65-70.

Mosely, E. C., Gorham, D. R., & Hill, E. Computer scoring of inkblot perceptions. Perceptual and Motor Skills, 1963, 17, 498.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. The General Inquirer: a computer approach to content analysis. Cambridge, Mass.: The MIT Press, 1966.

Swenson, W. M., Rome, H. P., Pearson, J. S., & Brannick, T. L. A totally automated psychological test. Journal of the American Medical Association, 1965, 191(11), 925-927.

Veldman, D. J., & Menaker, Shirley. Computer applications in assessment and counseling. Journal of School Psychology, 1968, 6, 167-176.

Veldman, D. J. Fortran programming for the behavioral sciences. New York: Holt, Rinehart & Winston, 1967.

Veldman, D. J. Computer-based sentence completion interviews. Journal of Counseling Psychology, 1967, 14(2), 153-157.