1984

The Supplemental Ability of "Thinking-Aloud" Data in the Psychometric Assessment of Aptitude Measures

Ann Reed Gaines
Loyola University Chicago

THE SUPPLEMENTAL ABILITY OF "THINKING-ALOUD" DATA IN THE PSYCHOMETRIC ASSESSMENT OF APTITUDE MEASURES

by

Ann Reed Gaines

A Dissertation Submitted to the Faculty of the Graduate School of Loyola University of Chicago in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

May

1984

ACKNOWLEDGMENTS

I would like to express my appreciation to the individuals who served as the members of my Dissertation Committee for their guidance and assistance during the preparation of this dissertation. The Dissertation Committee included Dr. Jack Kavanagh, Director, Dr. Judy Irwin, Dr. Steven Miller, and Dr. Ronald Morgan. I would further like to express my gratitude to the four individuals who graciously volunteered hours of their time to serve as the subjects for the present study and, unfortunately, due to the ethics of research, must remain anonymous.

VITA

The author, Ann Reed Gaines, was born in Detroit, Michigan on September 11, 1949. She attended primary school primarily in Pittsburgh, Pennsylvania and secondary school primarily in Wheaton, Illinois, graduating from Wheaton Central High School in 1967. She received a Bachelor of Science degree in Medical Technology from the University of Kentucky; Lexington, Kentucky in May, 1971. In December, 1974 she received a Master of Science degree in Education, likewise from the University of Kentucky. She began a Doctor of Philosophy degree in Education at Loyola University of Chicago; Chicago, Illinois in 1978.

She was employed as a medical technologist from 1971 to 1973 by International Clinical Laboratories; Lexington, Kentucky and as a assistant supervisor and medical technologist by Orange Memorial Hospital; Orlando, Florida from 1973 until 1974. She was the Assistant Educational Coordinator, School of Medical Technology, Good Samaritan Hospital and International Clinical Laboratories; Lexington, Kentucky from 1974 to 1976. From 1976 until 1982, she was the Educational Coordinator, Programs in Medical Technology and an Associate, Department of Pathology at Northwestern University Medical School; Chicago, Illinois. She pursued completion of her doctorate on a full-time basis from 1982 until 1983. Since 1983, she has been a Research Assistant, Health Pro-

fessions Education at the Center for Educational Development of the University of Illinois at Chicago; Chicago, Illinois.

She was appointed to the Research and Development Committee of the Board of Registry of the American Society of Clinical Pathologists in 1977 and served as a member of that committee until 1982. The Research and Development Committee addresses issues concerning the national certification examinations offered by the Board of Registry for non-physician clinical laboratory professions, issues relevant to continuing education for the clinical laboratory professions, among others.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

Table | Page

CHAPTER I

STATEMENT OF THE PROBLEM

Within the context of contemporary intelligence theory, aptitude measures are recognized to represent varying degrees of univariate and multivariate, linear as well as nonlinear, continuous as well as discontinuous, homogeneous and heterogeneous measures. Sources of variance in aptitude measures are acknowledged to include, but not be restricted to, attributes of the subjects, the measures, and/or circumstances of administration. Consequently, the sources of variance in aptitude measures cannot be presumed a priori to be invariant; cannot be presumed a priori to result in intrinsic score variance; and cannot, according to a specific paradigm or mathematical model, be partitioned a priori into mutually exclusive and exhaustive components (Snow, 1979; Bloom and Broder, 1950; Morrison, 1960; Nunnally, 1978; Hunt and MacLeod, 1979; Detterman, 1979; Humphreys, 1974, 1976; French, 1957, 1965; Pellegrino and Glaser, 1979; Kropp, Stoker, and Bashaw, 1966; Lerner, 1976; Sternberg, 1977; Bower and Hilgard, 1981). The conceptualization of aptitude measures, within the context of intelligence theory, is approximated by the following summary and indicates that aptitude measures must be considered as:

> ... stimulus complexes which can be described by parameters of the stimulus set. Tests differ with respect to such stimulus parameters as the instructions given to subjects, the amount of preliminary practice, the number of items, the complexity of items, the number and similarity of response choices, the amount of irrelevant and redundant information, the time-limit conditions, and many others. The results of measurement depend upon the interaction of individual differences with such dimensions of the measurement situation (Morrison, 1960, pp. 232-233).

This conceptualization has implications regarding assessment of the validity and reliability of aptitude measures. The validity and reliability of aptitude measures are traditionally assessed within the context of measurement or psychometric theory. Correspondingly, validity and reliability are referenced to the mathematical model of linear regression and are expressed quantitatively as descriptive coefficients and/or inferential statistics. Assumptions underlying the psychometric assessment of validity and reliability include that aptitude measures represent univariate, linear, continuous, and homogeneous measures. The sources of variance in aptitude measures are considered invariant and are partitioned into true and error variance components, attributable to interindividual differences in the level of aptitude(s) and random errors of measurement, respectively. Interpretation of psychometric coefficients and statistics is predicated on intrinsic score variance in aptitude measures (Hays, 1973; Nunnally, 1978; Popham, 1978; Thorndike and Hagen, 1977; Edwards, 1976; Kerlinger, 1973).

The somewhat disparate conceptualizations of aptitude

measures, within the respective contexts of intelligence theory and psychometric theory, suggest that exclusive reliance on the psychometric assessment of aptitude measure validity and reliability may not be appropriate and warranted in all instances. Suggested is that for some aptitude measures, the assumptions underlying the psychometric assessment of validity and reliability may be violated or, more importantly, may not consider all relevant sources of variance and may not adequately partition all sources of variance in a relevant manner. For those measures where psychometric assessment is neither appropriate nor warranted, suggested is that the traditional descriptive and inferential interpretations of psychometric coefficients and statistics may correspondingly be inappropriate and unwarranted.

Seemingly what is needed is a means of providing supplemental data to that utilized in the psychometric assessment of aptitude measure validity and reliability. Supplemental data could be utilized to indicate whether or not the assumptions underlying psychometric assessment, the partitioning of variance in psychometric assessment, and the traditional descriptive and inferential interpretations of psychometric coefficients and statistics are appropriate and warranted. If not, supplemental data could be utilized to enhance the descriptive and inferential interpretations of validity and reliability coefficients and statistics, by suggesting relevant limitations or qualifications for the

interpretations. Given that the sources of variance in aptitude measures may include, among others, attributes of the subjects, the measures, and/or circumstances of administration, seemingly what is further needed is a means of providing supplemental data at the level of subjects, measures, and circumstances of administration, at a minimum.

One type of supplemental data to that utilized in the psychometric assessment of aptitude measure validity and reliability which explicitly or implicitly considers subjects, measures, and circumstances of administration is "thinking-aloud" data. Thinking-aloud data, by definition, consist of the verbalized responses of single subjects obtained concurrently with the individual administration of single item measures. Evidence from the literature suggests that thinking-aloud data can provide supplemental data relevant to the psychometric assessment of aptitude measure validity and reliability (e.g., multivariate measures, intra-individual differences or discontinuities). Evidence from the literature further suggests that thinking-aloud data can be shown to possess both internal and external validity (Bloom and Broder, 1950; Lieberman, 1979; Newell and Simon, 1972; Olshavsky, 1976-1977; Fareed, 1971; Kavale and Schreiner, 1979; Bower and Hilgard, 1981).

The purpose of the present study was to assess the supplemental ability of thinking-aloud data in the psychometric evaluation of aptitude item validity and reliability. Twen-

ty-five items of the types generally included on standardized aptitude examinations were individually administered to four subjects by means of a nonschedule standardized interview developed for the present study. Both items and subjects were selected by means of matrix sampling strategies. The nonschedule standardized interview was utilized to elicit the thinking-aloud responses of the subjects to the items and to various aspects of the items and subjects' responses to the items. Transcripts of the thinking-aloud responses constituted the data base for the present study.

The transcripts were content analyzed to derive what were termed psychometric inferences, or inferences relevant to the validity and reliability of the items. Three types of psychometric inferences were derived and were designated content/construct validity, internal consistency/discrimination, and alternate form/test-retest reliability. The psychometric inferences for each item were compared to the psychometric data available for each item to assess the extent to which the psychometric inferences supplemented the psychometric data. The psychometric data for each item were restricted to the operational definition of the aptitude purported to be measured by the item, as no other psychometric data (e.g., item analysis indices) were available or obtainable. Further content analysis of the transcripts and within-method and between-method triangulation were utilized to derive what were termed methodological inferences, or in-

ferences relevant to the internal and external validity of the three principal components of the present study: the subjects as the data sources, the nonschedule standardized interview as the means of data collection, and the investigator as the content analyst.

The present study was formulated within the context of exploratory methodological research in psychometrics and was conducted by means of a qualitative research paradigm. In contrast to the traditional quantitative research paradigm, no independent or dependent variables were specified, and no statistical hypotheses were declared. A restatement of the purpose of the present study constituted the research question: To what extent do thinking-aloud data provide supplemental data relevant to the psychometric assessment of aptitude item validity and reliability?

CHAPTER II

REVIEW OF THE LITERATURE

The purpose of the review of the literature is to support the premises on which the present study is based. Specifically, for some aptitude measures, the assumptions underlying psychometric assessment, the partitioning of variance in psychometric assessment, and the traditional descriptive and inferential interpretations of psychometric assessment may be inappropriate and unwarranted. The purpose of the review of the literature is further to support the rationale underlying the present study. Specifically, supplemental data to that considered and utilized in the psychometric assessment of aptitude measure validity and reliability enhance the descriptive and inferential interpretations of psychometric coefficients by suggesting relevant limitations or qualifications for the interpretations.

Two means of accomplishing these purposes are utilized. First, the results of studies or other findings are provided in which supplemental data to that utilized in psychometric assessment of aptitude measures enhance the interpretation of validity and reliability coefficients. For example, among others, instances are cited in which the score variance in aptitude measures is attributable to other sources

of variance (e.g., strategies, "practice") as well as to interindividual differences in the level of the aptitude purported to be measured (e.g., Hunt and MacLeod, 1979; Swinton and Powers, 1983). Second, the results of studies or other findings are provided in which supplemental data to that utilized in psychometric assessment of aptitude measures are needed to enhance the interpretation of validity and reliability coefficients. For example, among others, instances are cited in which more than one traditional interpretation of validity and reliability coefficients is possible and in which the possible interpretations are somewhat disparate, due to the lack of data beyond that considered and utilized in the psychometric assessment of validity and reliability (e.g., Ekstrom, French, and Harman, 1976b; Sternberg, 1977).

Of necessity, the results of studies or other findings cited represent a survey (i.e., breadth), rather than an exhaustive summary (i.e., depth) of the literature. The instances cited are purposively selected to illustrate various and diversified aspects of the premises and rationale underlying the present study. By virtue of the fact that each of the instances provided may illustrate more than one aspect of the premises and rationale of the present study, each instance cited is presented in a separate section. In conjunction with this fact, each section is labeled only by means of the source on which the content of the section is based (i.e., a section heading of Hunt and MacLeod, 1979),

as titles or headings which concisely indicate or summarize the content of each section are not devisable.

Bloom and Broder, 1950

Eight students, ranging in age from 15 to 25 and placing at or above the fiftieth percentile on an unspecified standardized examination norm-referenced for college freshmen, were individually administered various vocabulary items. Content analysis of the thinking-aloud responses of the subjects to the vocabulary items revealed that the subjects utilized various word-related strategies for words which were unfamiliar, as follows:

> Thus, [for the word portent], several of the students decided that portent sounded like a noun and that omen was the only other noun which could apply. They ruled out mobile and conceited on the grounds that these were not similar parts of speech (p. 65).
>
> Thus, [for the word anomalous], several of the students decided that nom in anomalous referred to name and that a referred to without. These students then selected nameless as the synonym. Although this was a perfectly good method of problem-solving, it did not help these students in finding the correct response - irregular. This technique, however, did aid several of the students in getting the correct synonym for corpulent ... . Here they related corpus to the Latin for body, then selected portly as the most appropriate term to apply to body (p. 65).

Neither the frequency with which subjects utilized these and similar strategies nor the proportion of successful and unsuccessful applications of strategies was reported.

The results of this study provide supplemental data relevant to the validity and reliability of vocabulary items.

That word-related strategies may constitute a source of variance is indicated, although the extent to which strategies may constitute a systematic source of variance in vocabulary scores is indeterminate. Further indicated is that word-related strategies, as a source of variance, may be discontinuous as well as not invariant within and between subjects and items. The results of this study are in contrast and supplemental to previous conceptualizations of vocabulary measures of verbal comprehension as univariate measures, with variance attributable only to interindividual differences in the level of vocabulary (Nunnally, 1978; Guilford, 1967).

Ekstrom et al., 1976b

Two examination measures of a factor termed figural flexibility were pretested with from 625 to 746 male naval recruits. The mean scores of the subjects on the two examinations were 6.1 and 1.3. Following an unspecified revision in the directions for the examinations, the measures were posttested with from 542 to 574 male naval recruits, described as "similar but probably less able" (p. 7). The mean scores of these subjects on the two examinations were approximately 8.2 and 2.0, respectively. The difference in the mean scores for the pre- and post-revision administrations of the examinations was interpreted as reflecting "obviously a major change in test difficulty (apparently the

revised directions made these tests much simpler)" (p. 7). No standard deviations, reliability coefficients, or validity coefficients were provided for the pre- and post-revision examination scores.

The results of this study provide supplemental data relevant to the validity and reliability of the figural flexibility measures. That the directions provided for the examinations constituted a source of variance in examination scores is indicated. However, supplemental data are needed concerning the type of revision in the directions, in order to determine what specific confounding influence existed in the pre-revision directions. The manner in which the revised directions reduced or eliminated the confounding influence is needed to assess the extent to which the revised directions may or may not have systematically affected the validity and reliability of the measures. The manner in which the revised directions reduced or eliminated the confounding influence is needed to assess whether or not the confounding influence was invariant within and between subjects as well as items. The types of supplemental data needed include, yet are probably not restricted to, standard deviations, reliability coefficients, validity coefficients, and thinking-aloud responses of the subjects for the pre- and post-revision examinations.

Hunt and MacLeod, 1979

Items of the type variously referred to as sentence-picture verification or sentence-picture comparison were administered to 59 college students. Variable measures obtained for each subject included reaction times for responding to the sentence-picture items, scores on unspecified verbal and spatial ability measures, and whether subjects represented the sentence-picture stimulus (e.g., $\overset{+}{*}$) in memory in a semantic medium (e.g., the "plus" is above the "star") or in a figural medium (e.g., $\overset{+}{*}$).

For subjects utilizing a semantic representation of the stimulus (n = 43), the partial correlation coefficient between reaction time and verbal ability scores, with the effect of spatial ability removed, was $r = -.44$, $p < .01$; the partial correlation coefficient between reaction time and spatial ability scores, with the effect of verbal ability removed, was $r = .07$, NS (not significant). For subjects utilizing a figural representation of the stimulus (n = 16), the exact reverse relationship was manifested. The partial correlation coefficient between reaction time and verbal ability scores, with the effect of spatial ability removed, was $r = -.05$, NS; the partial correlation coefficient between reaction time and spatial ability scores, with the effect of verbal ability removed, was $r = -.64$, $p < .01$. The mean reaction time for responding to the sentence-picture items, interpolated from a graph, was 1200 "units" for sub-

jects utilizing a semantic representation and 650 units for subjects utilizing a figural representation. Certain subjects, although the proportion was not specified, were capable of utilizing either a semantic or a figural representation of the sentence-picture stimulus. Neither the partial correlation coefficients nor the mean reaction time was reported for the composite sample of subjects.

The results of this study provide supplemental data relevant to various aspects of the validity and reliability of the sentence-picture comparison items. In terms of content validity, the figural medium in which the sentence-picture items were depicted did not invariably correspond to the medium in which subjects "processed" the items (i.e., figural, semantic, figural and/or semantic). Such a premise has traditionally been the basis underlying utilization of figural or symbolic media for so-called "culture-free" aptitude measures; that is, that figural and/or symbolic media remove the semantic constraints of items for "disadvantaged" subjects (Reynolds and Jensen, 1983; Brody and Brody, 1976; Butcher, 1970). In terms of construct validity, whether the items measured primarily an aptitude analogous to verbal ability or analogous to spatial ability depended upon the medium in which subjects represented the stimulus in memory; that is, at a minimum, the sentence-picture comparison items constituted varying degrees of bivariate aptitude measures. The aptitude(s) measured by the items were discontinuous be-

tween subjects (i.e., verbal versus spatial) and were not invariant within all subjects (i.e., those capable of alternating the medium in which the stimulus was represented). In conjunction with the medium in which the stimulus was represented, the level of spatial ability, and the level of verbal ability, reaction time was a source of variance in the items. However, by virtue of the fact that the sentence-picture items were of what has been termed "trivial difficulty", or capable of being responded to correctly in the absence of restrictive time limits allowed for administration (Nunnally, 1978; Guilford, 1967; Morrison, 1960), the effect of these sources of variance as determinants of item scores is indeterminate. Suggested is that under restrictive time limits allowed for administration, reaction time (i.e., "speededness") would constitute a source of variance extraneous to interindividual differences in the levels of aptitude, unless reaction time constituted an essential component or aptitude and was specified in an operational definition for sentence-picture comparison items. "Speededness" has been specified as an essential component of other aptitudes (Nunnally, 1978; Guilford, 1967; Ekstrom et al., 1976b; Tyler, 1979). Thus, had supplemental data, in the form of the medium in which subjects represented the sentence-picture stimulus in memory, not been provided, sentence-picture items might have been presumed to measure some other aptitude(s) (e.g., perhaps perceptual speed) rather than apti-

tudes analogous to verbal and spatial abilities. Further, sentence-picture items might have been presumed to measure verbal and spatial abilities somewhat comparably across subjects. Had supplemental data, in the form of reaction time variables, not been provided, the interactive effect of this source of variance with the medium in which subjects represented the stimulus, verbal ability, and spatial ability might not have been discerned.

Sternberg, 1977

The scores of 16 college students were obtained on the following aptitude measures: 60 verbal analogy items selected from the information bulletin distributed by the publisher of the Miller Analogies Test (MAT), three so-called "reference ability" reasoning examinations, four so-called reference ability vocabulary examinations, and 30 items described as animal name analogies (e.g., gorilla is to deer as bear is to [cow, pig, tiger, or monkey]). The correlation coefficients between the MAT scores and scores of the other measures were as follows: reasoning, $\underline{r} = .77$, $\underline{p} < .001$; vocabulary, $\underline{r} = .76$, $\underline{p} < .001$; animal name analogies, $\underline{r} = .34$, NS. The partial correlation coefficient between the MAT and reasoning scores, with the effect of vocabulary removed, was $\underline{r} = .64$, $\underline{p} < .01$. The partial correlation coefficient between MAT and vocabulary scores, with the effect of reasoning removed, was not reported. An operational definition,

as such, for the MAT was that the MAT measured "... scholastic aptitude at the graduate school level. .... The test items require the recognition of relationships rather than display of enormous erudition" (p. 301). Operational definitions for the other items and examinations utilized were not provided.

The results of this study were interpreted as supporting the MAT as a measure of reasoning ability, although not exclusive of vocabulary, given the correlation coefficients between the MAT and reasoning scores, the MAT and vocabulary scores, and the MAT and reasoning scores with the effect of vocabulary removed. Interpretation of the not significant correlation coefficient between the MAT and animal name analogy scores was as follows: "The low correlation between the animal name and Miller analogies is probably due to lack of overlapping variance in both reasoning and vocabulary (p. 307).

Analogy items, regardless of the type, have traditionally been considered to constitute measures of inductive reasoning (Green, Guilford, Christensen, and Comrey, 1953; Nunnally, 1978; French, 1957, 1965; Sternberg, 1977). The dismissal of a not significant correlation coefficient between the MAT and animal name analogies scores as attributable to "lack of overlapping variance in both reasoning and vocabulary", without further elaboration, seemingly constitutes a cavalier interpretation. That is, unless supple-

mental data indicates that animal name analogy items are invalid measures of inductive reasoning, are devoid of semantic content, and/or are unreliable measures, the construct validity interpretation of animal name analogy items as lacking reasoning and vocabulary components is inappropriate and unwarranted.

If the lack of a statistically significant correlation coefficient between the MAT and animal name analogy items is, in fact, attributable to "lack of overlapping variance in both reasoning and vocabulary", supplemental data are needed to interpret the correlation coefficients between the MAT scores and those of the three so-called reference ability measures of reasoning, designated as word grouping, letter series, and Cattell reasoning, but not described. Presumably, the word grouping items were of the type in which four or five words are presented as a group and in which subjects are to determine "which word does not belong with the others?". The correlation coefficient between the MAT and word grouping scores was $\underline{r} = .66$, $\underline{p} < .01$, presumably reflecting common variance attributable to both reasoning and vocabulary. Presumably, the letter series items were of the type in which various numbers of letters are presented, in which one letter of the series has been omitted, and in which subjects are to determine what letter has been omitted from the series. The correlation coefficient between the MAT and letter series scores was $\underline{r} = .72$, $\underline{p} < .01$, presum-

ably reflecting common variance attributable to reasoning only. Supplemental data are needed to facilitate interpretation of why the word grouping and letter series correlation coefficients are of comparable magnitude. Within the context of convergent and discriminant or multitrait-multimethod construct validity (Kerlinger, 1973), these correlation coefficients are ambiguous in terms of the construct validity of the MAT. Supplemental data are further needed to facilitate interpretation of what common variance is reflected by the correlation coefficient between the MAT and Cattell reasoning scores; no information concerning what types of items are included in that measure is provided or can be presumed.

Supplemental data are needed to enhance interpretation of the results of this study. The types of supplemental data needed include, yet probably are not restricted to: score means and standard deviations, reliability coefficients, and validity coefficients for the various measures; the partial correlation coefficient between MAT and vocabulary scores, with the effect of reasoning removed; operational definitions for the various measures; and perhaps thinking-aloud responses of the subjects to the animal name analogy items particularly, to facilitate interpretation of if and why such items lack both reasoning and vocabulary as sources of variance.

Ekstrom et al., 1976b

Two presumably alternate form examinations are included as measures of verbal comprehension in the Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976a). The first examination consists of four-option multiple-choice vocabulary items in which the options are labeled by means of Arabic numberals and are arranged in a horizontal array. The directions for the examination state that the response for each item is to be indicated by writing, in a set of parentheses placed at the far right of the array of options, the number corresponding to the option selected. The second examination consists of five-option multiple-choice vocabulary items in which the options are labeled by means of Arabic numerals but are arranged in a vertical array. The directions for the examination state that the response for each item is to be indicated by drawing an "X" through the number corresponding to the option selected. In the description provided for these two examinations in the manual accompanying the Kit of Factor-Referenced Cognitive Tests is the statement that "[t]he format [of the second examination] is intentionally different from that of [the first examination] to reduce common factor variance of an artifactual nature" (p. 164). No elaboration of this statement is provided. Reliability coefficients of $\underline{r}$ = .70 and $\underline{r}$ = .68 are reported for the first and second examinations, respectively, for a sample of 294 sixth grade students. Although not re-

ported, given the n's and the r's, p < .001 (Downie and Heath, 1970).

In conjunction with variance presumably attributable to interindividual differences in the level of vocabulary measures of verbal comprehension, variance in one or both of these examinations is presumably attributable to the response formats of the items, or to what has been termed "bias" variance (Humphreys, 1974, 1976; Nunnally, 1978). Supplemental data are needed to facilitate interpretation of the reliability coefficients for this study, specifically concerning the type of "artifactual" variance contributed by the item response formats, the extent to and manner in which the artifactual variance interacts with interindividual differences in the level of the verbal comprehension aptitude purported to be measured, and the degree to which the response format alterations control for the artifactual variance. Further, the effect of the artifactual variance on the validity of the measures, if any, is needed. Presuming that artifactual variance of the type apparently present in these examinations is not unique to these examinations or to only vocabulary examinations, supplemental data such as that specified above are needed to assess the implications of multiple-choice response formats on the validity and reliability of other aptitude measures.

Bloom and Broder, 1950

Eight students, ranging in age from 15 to 25 and placing at or above the fiftieth percentile on an unspecified standardized examination norm-referenced for college freshmen, were individually administered, among others, a "geology" and an "algebra" examination item. For the geology item, subjects were to "[r]ank the following life forms in the order of their appearance in the geologic record" (p. 45). Content analysis of the thinking-aloud responses of the subjects for this item revealed that:

> Students were confused as to whether [the directions] referred to the oldest or the most recent life forms, since the 'order of appearance' might refer to the order in which they appeared chronologically or to the order in which they would appear as the geologic record is uncovered (p. 45).

For the algebra item, the multiple-choice options provided were presented as follows:

"A- $x = 3y$, B- $x = y^3$, C- $xy - 3$, D- $x + y = 3$, E- $\frac{x}{2} = \frac{y}{3}$" (p. 44), and content analysis of the thinking-aloud responses of the subjects for this item revealed that:

> Some of the students read the alternatives as 'A minus x equals 3y', 'B minus x equals y [cubed]', etc. This, of course made a problem which was impossible to solve. Frequently the student would recognize and correct the error after he attempted to solve the problem and found that it made no sense (p. 44).

No information was provided concerning the proportion of subjects having misunderstood either the directions or the options for the two items, respectively. Neither was any information analogous to item analysis indices (i.e., diffi-

culty, discrimination) provided.

The supplemental data provided by means of the thinking-aloud responses of the subjects to both items are relevant to the validity and reliability of the items. First, the supplemental data support that supplemental data are needed at both the level of subjects and items, rather than only at the level of samples of subjects and items (i.e., examinations). Second, the supplemental data provided by the thinking-aloud responses indicate that for subjects who misunderstood either the directions or the options for the respective items, neither item constituted a valid and discriminating measure of the geology and algebra aptitude(s) presumed to be measured. Had item analysis data for both items been available and indicated "good" items, the traditional interpretations of such indices would have been inappropriate and unwarranted. Had item analysis data for both items been available and indicated "poor" items, the specific attributes of the items which may have contributed to the poor item analysis indices may not have been discerned and may have, instead, been interpreted within the context of the levels of geology and algebra knowledge of the subjects.

Swinton and Powers, 1983

An experimental group of college students (n = 25) received seven contact hours of instruction described as:

> ... focusing on strategies and techniques specific to the analytical portion of the GRE [Graduate Record Examination] Aptitude Test and to its specific item formats rather than on development of the cognitive abilities that the test is designed to measure (p. 406).

The control group (n = 415) received no such instruction. Mean analytical ability scores, expressed in terms of standardized scores which may range from 200 to 800, for the two groups of subjects from an actual administration of the GRE Aptitude Test were 530.7 for the control group and 591.8 for the experimental group. The difference between the mean scores for the two groups of subjects was statistically significant at the .05 level. Three types of items were represented in the analytical ability section of the GRE Aptitude Test. For the first type of item (n = 40), termed analysis of explanations, the difference between the 24.2 mean score for the control group and the 28.6 mean score for the experimental group was statistically significant at the .001 level. For the second type of item (n = 15), termed logical diagrams, the difference between the 10.7 mean score for the control group and the 12.1 mean score for the experimental group was statistically significant at the .05 level. For the third type of item (n = 15), termed analytical reasoning, and in actuality, containing two types of items, the difference between the 7.2 mean score for the control group and the 7.5 mean score for the experimental group was not statistically significant. The results of this study were interpreted as follows:

> In summary, it appears that scores of the analytical section of the GRE Aptitude Test, as constituted at the time of this study, may be improved under at least some conditions by relatively short-term interventions that focus primarily on practice and familiarization (p. 409).

Standard deviations and reliability coefficients for the item type subtests and the composite analytical section were not provided. Neither were operational definitions for the item types provided, although each type of item was briefly described, nor were any other indicants of the validity of the three types of items provided.

The results of this study provide supplemental data relevant to the construct validity of the analytical section of the GRE Aptitude Test. The results suggest that, in conjunction with interindividual differences in the analytical aptitude(s) purported to be measured by the analysis of explanation and logical diagrams items, interindividual differences in the "... facility with the methods of assessment or familiarity with the format of items" (p. 404) may likewise constitute a source of variance. Supplemental data are needed concerning the extent to which the two types of analytical reasoning items are "homogeneous", given that inspection of the two types of analytical reasoning items suggests that one type more closely resembles the analysis of explanation items than the second type of analytical reasoning items. Thus, the possibility that the lack of a statistically significant difference between the control and experimental groups of subjects on the analytical reasoning

is attributable to having not partitioned each of the two types of analytical reasoning items into separate categories for analysis. Separate partitioning of the two types of analytical reasoning items would have resulted in an extremely small sample size of each type of item, however, as only 15 items constituted the analytical reasoning section of the GRE Aptitude Test.

French, 1957

An inductive reasoning measure, variously termed letter sets or letter groups, was administered to 361 military academy freshmen. Each of the two parts of the examination consisted of 15 items and was allotted an administration time limit of five minutes. Reported was that only seven per cent of the subjects completed the first part and that only 20 per cent of the subjects completed the second part. The reliability coefficient for the examination, in the form of alternate form reliability between the two parts, was $\underline{r}$ = .43. Although not reported, given the $\underline{n}$ and the $\underline{r}$, $p < .001$ (Downie and Heath, 1970). The basis on which subjects were determined to have completed both parts of the examination was not reported. Neither were the mean scores nor standard deviations for the two parts or for the composite examination reported.

Based only on the reported reliability coefficient, three traditional descriptive and inferential interpreta-

tions of the reliability coefficient are suggested (Edwards, 1976; Nunnally, 1978; Downie and Heath, 1970; Kerlinger, 1973; Thorndike and Hagen, 1977). In terms of alternate form reliability, given $\underline{r}$ and $\underline{p}$, the first and second parts of the examination can be considered to have equivalently sampled items. With respect to construct validity and variance, the coefficient of determination (i.e., $\underline{r}^2 = .43^2 = .18$) indicates that only 18 per cent of the variance between the two parts of the examination is accounted for by inductive reasoning. With respect to construct validity and variance, the coefficient of nondetermination (i.e., $1 - r^2 = 1 - .18 = .82$), indicates that 82 per cent of the variance between the two parts of the examination is unaccounted for by inductive reasoning. However, to what source(s) of variance the 82 per cent is attributable is indeterminate. The first interpretation of the reliability coefficient indicates that the letter sets measure is relatively reliable. The second interpretation implies that the measure is relatively invalid. The third interpretation implies that the measure is invalid and/or unreliable. Supplemental data are needed to enhance interpretation of the psychometric coefficients for this examination.

Reliability coefficients for a similar, if not identical, examination administered with a five-minute time limit to seemingly comparable subjects range from .74 to .84 (Ekstrom et al., 1976b). Suggested from these results is

that the reliability coefficient of .43, although statistically significant, may be relatively low due to the administration time limit of seven minutes (i.e., "speededness"). Had the means and standard deviations of scores on the letter sets examinations been available and compared across the samples of subjects (i.e., French, 1957; Ekstrom et al., 1976b), the relatively low reliability coefficient in the former study might be suggested to be attributable to a lack of score variance or so-called "restriction of range". As only seven and 20 per cent of the subjects completed the first and second parts, respectively, of the examination, the reliability coefficient of .43 may have been based on a relatively considerable amount of "missing" data. That is, for an unknown proportion of the 15 items in each part of the examination, psychometric coefficients may have been based on as few as 26 subjects' responses (i.e., 7 per cent of 361 subjects) and 72 subjects' responses (i.e., 20 per cent of 361 subjects). The manner in which subjects responded to the items of the examination may likewise have resulted in the psychometric coefficients having been based on a biased and nonrandom subsample of subjects (e.g., dependent upon whether subjects responded to all attempted items, "almost" completed the examination, and/or responded to only the "easy" items. Supplemental data relevant to the considerations delineated would enhance interpretation of the psychometric coefficients for this study, as all such

considerations have been associated with spurious correlation coefficients (Hays, 1973; Edwards, 1976; Nunnally, 1978; Kerlinger, 1973; Nie, Hull, Jenkins, Steinbrenner, and Bent, 1975).

The purpose of the review of the literature was to support the premises on which the present study was based. Specifically, for some aptitude measures, the assumptions underlying psychometric assessment, the partitioning of variance in psychometric assessment, and the traditional descriptive and inferential interpretations of psychometric assessment may be inappropriate and unwarranted. The purpose of the review of the literature was further to support the rationale underlying the present study. Specifically, supplemental data to that considered and utilized in the psychometric assessment of aptitude measure validity and reliability enhance the descriptive and inferential interpretations of psychometric coefficients by suggesting relevant limitations or qualifications for the interpretations. The results of studies or other findings were provided in an effort to accomplish these purposes.

CHAPTER III

METHODOLOGY

The purpose of the present study was to assess the supplemental ability of thinking-aloud data in the psychometric evaluation of aptitude item validity and reliability. To do so, twenty-five items of the types generally included on standardized aptitude examinations were individually administered to four subjects by means of a nonschedule standardized interview developed for the present study. Both items and subjects were selected by means of matrix sampling strategies. The nonschedule standardized interview was utilized to elicit the thinking-aloud responses of the subjects to the items and to various aspects of the items and subjects' responses to the items. Transcripts of the thinking-aloud responses constituted the data base for the present study.

The transcripts were content analyzed to derive what were termed psychometric inferences, or inferences relevant to the validity and reliability of the items. Three types of psychometric inferences were derived and were designated content/construct validity, internal consistency/discrimination, and alternate form/test-retest reliability. The psychometric inferences for each item were compared to the

psychometric data available for each item to assess the extent to which the psychometric inferences supplemented the psychometric data. The psychometric data for each item were restricted to the operational definition of the aptitude purported to be measured by the item, as no other psychometric data (e.g., item analysis indices) were available or obtainable. Further content analysis of the transcripts and within-method and between-method triangulation were utilized to derive what were termed methodological inferences, or inferences relevant to the internal and external validity of the three principal components of the present study: the subjects as the data sources, the nonschedule standardized interview as the means of data collection, and the investigator as the content analyst.

The aspects of the present study which collectively constituted the methodology were as follows:

- the nonschedule standardized interview,
- the sample of aptitude examination items,
- the sample of subjects,
- the procedure, and
- the content analysis.

Each of these aspects is detailed in the following sections.

Nonschedule Standardized Interview

A nonschedule standardized interview was adopted as the means of data collection for the present study for three reasons. First, a nonschedule standardized interview provided a means for eliciting the thinking-aloud responses of

the subjects to the items and to various aspects of the items and subjects' responses to the items. Second, a nonschedule standardized interview enabled the types of supplemental data sought in the thinking-aloud responses to be specified a priori, thus ensuring that comparable data would be obtained across all subjects and items. Third, a nonschedule standardized interview accorded sufficient flexibility that the sequence of the inquiries posed to subjects could be varied, if necessary, and that the responses of the subjects could be pursued by the investigator in greater depth, if deemed relevant, to provide additional information, clarification, or other elaboration (Denzin, 1978; Patton, 1980; Kerlinger, 1973).

The nonschedule standardized interview consisted of two basic sections, intended to elicit from the subjects two general types of data. The first section was intended to elicit the responses of the subjects to the items as aptitude measures. The content of this section was suggested by the literature relevant to task analytic approaches to the study of aptitudes or intelligence. Based on the task analytic research in this area, responding to an aptitude item measure proceeds through specific phases (Bower and Hilgard, 1981; Sternberg, 1977; Fleishman, 1975); paralleling the phases suggested by such task analytic research, subjects were requested to:

- read aloud both the directions for the item and the

item;
- describe aloud "what" was perceived to be required for responding to the item and the manner in which responding to the item would be approached;
- respond aloud to the item; and
- describe aloud the means by which closure was achieved on the response generated or selected for the item (e.g., for a multiple-choice item, the manner in which incorrect options had been eliminated from further consideration), if not explicitly or implicitly stated while responding aloud to the item.

The second section of the nonschedule standardized interview was intended to elicit the responses of the subjects to various aspects of the items and subjects' responses to the items. The content of this section was suggested, in part, on the studies and other findings presented in the Review of the Literature chapter and, in part, on a rational, subjective basis. After responding aloud to the item, subjects were asked:

- "what" they perceived the item to have measured (e.g., abilities, knowledge);
- what other approaches they could have utilized for responding to the item;
- how they would "double-check" their item response;
- whether they had had previous exposure to or experience with the general type of item;
- if so, to what extent was prior familiarity with the general type of item an asset in responding to the item;
- whether responding to the item approximated any activity engaged in by them on a somewhat routine basis (e.g., in work-related contexts, in "hobby"-related contexts);
- whether the item was "easy" or "difficult" and for what reason(s); and
- whether there were any additional, miscellaneous comments or remarks concerning any aspect of the item or their responses to the item.

The nonschedule standardized interview was pretested, as such, in various informal pilot studies conducted prior to and in preparation for the present study. Subjects and

items utilized in the pilot studies were comparable to those utilized in the present study. The phrasing of the various inquiries of the nonschedule standardized interview was considered sufficiently revised and refined when the responses of the subjects in the pilot studies approximated anticipated responses; the phrasing was considered sufficiently "uncued" by virtue of the fact that subjects' responses were diversified and not stereotypical.

Aptitude Examination Items

A matrix sampling strategy was utilized to select the aptitude items for the present study as a means of enhancing the objectivity and randomness of item selection. The variables incorporated into the matrix sampling strategy were intended to constitute only attribute variables of the items, rather than independent or dependent variables. The attribute variables of items which might affect the validity and reliability of the items were suggested by the literature and were restricted to those which could be determined or classified on a rational, objective basis by the investigator, as follows:

- the factor, within a factor analytic context, purported to be measured by the item (e.g., Ekstrom et al., 1976b; Butcher, 1970; Huttenlocher, 1976; Kaufman, 1981; French, 1957, 1965; Mukherjee, 1975; Naglieri, Kaufman, and Harrison, 1981; Nunnally, 1978; Green et al., 1953; Pellegrino and Glaser, 1979; Kropp and Stoker, 1966; Brody and Brody, 1976);
- the cognitive processes (e.g., categories of Bloom's taxonomy) presumed to be elicited by the item (e.g., Bloom, 1956; Kropp and Stoker, 1966; Guilford, 1967;

Kropp et al., 1966; Seddon, 1978; Poole, 1971);
- the content or medium (e.g., semantic, symbolic, figural) in which the item was expressed (e.g., Guilford, 1967; Hunt and MacLeod, 1979; Pellegrino and Glaser, 1979; Butcher, 1970; Brody and Brody, 1976; Reynolds and Jensen, 1983; Mukherjee, 1975; Kaufman, 1981);
- the response format (e.g., selected response, constructed response) in which the item was posed (e.g., Popham, 1978; Pellegrino and Glaser, 1979; Kropp and Stoker, 1966; Bloom and Broder, 1950; Swinton and Powers, 1983; Thorndike and Hagen, 1977; Nunnally, 1978);
- the possibility that response strategies would be elicited from the subjects by the item (e.g., Carroll, 1976; Pellegrino and Glaser, 1979; Ekstrom et al., 1976b; Kropp and Stoker, 1966; Bloom and Broder, 1950; French, 1965; Guilford, 1967; Educational Testing Service, 1982); and
- (the availability of a task analysis for the item, however, for another purpose; see the Content Analysis section of this chapter).

On the basis of informal pilot studies conducted prior to and in preparation for the present study, the maximum sample size of items feasible was determined to be 25. The distribution of the sample of items in terms of the matrix sampling strategy attribute variables is presented in Table 1, with one exception. The cognitive processes elicited by the items had been estimated based on the responses of only one subject during an informal pilot study. The distribution of the items in terms of the cognitive processes utilized (i.e., Bloom's taxonomy) represented, at most, an approximation and is summarized only as follows. Each of the six taxonomic categories specified in Bloom's taxonomy was presumed to be represented by a minimum of three items, with the exception of the "synthesis" category, which was not represented, as no relevant items were located. Further,

Table 1

Aptitude Examination Items: Distribution of Matrix Sampling Strategy Attribute Variables

Factor	n
Verbal	
Comprehension	2
Fluency	3
Reasoning	
General	3
Deductive	7
Inductive	4
Memory	
Rote	1
Spatial	
Visualization	1
Perceptual	
Speed	1
Flexibility of closure	1
Miscellaneous	
Integrative processes	1
Flexibility of use	1
Content	
Semantic	9
Symbolic	12
Figural	4
Format	
Selected response	
Exhaustive options	5
Nonexhaustive options	12
Constructed response	
Unrestrictive stipulations	5
Restrictive stipulations	3
Strategies	
Possible	13
Undocumented	12
Task analysis	
Available	6
Undocumented	19

the response format attribute variable had been subdivided to include both exhaustive (e.g., "none of the above") and nonexhaustive options for selected response items and unrestrictive and restrictive stipulations for constructed response items.

By virtue of the diversity of items needed to fulfill the attribute variables specified for the matrix sampling strategy, no one source examination could be located that contained items representative of all attribute variables. Therefore, items were selected from two standardized aptitude measures, the Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976a) and the Graduate Record Examination (Educational Testing Service, 1982). The only psychometric data relevant at the level of items provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) and the GRE 1982-83 Information Bulletin (Educational Testing Service, 1982) were the operational definitions of the aptitude(s) purported to be measured by the items. No further psychometric data (e.g., item analysis indices) were available from the authors or publisher of either aptitude measure (R.B. Ekstrom, Educational Testing Service, personal communication, November 19, 1982). The licensing agreement signed with the publisher of both measures prohibited the reproduction of the items, except for administration to the subjects in the present study, however, brief descriptive summaries of the items are presented

in the Results and Discussion chapter. The items selected by means of the matrix sampling strategy are identified by source in Table 2.

Subjects

A matrix sampling strategy was utilized to select the subjects for the present study, likewise as a means of enhancing the objectivity and randomness of subject selection. The variables incorporated into the matrix sampling strategy were intended to constitute only attribute variables of the subjects, rather than independent or dependent variables. The attribute variables of subjects which might affect the validity and reliability of the items were suggested by the literature (Brody and Brody, 1976; Butcher, 1970; Sternberg, 1974; McGrath, 1982; Huttenlocher, 1976; Dailey, 1959) and were restricted to those which could be ascertained by means of demographic inquiries to the subjects. The attribute variables utilized in the matrix sampling strategy included age, sex, educational level, and academic/occupational discipline. The first three attribute variables were dichotomized (i.e., 35-40, 51-56 years of age; male, female; baccalaureate, graduate and/or medical degree; respectively). As homogeneous as possible an academic/occupational discipline was utilized (i.e., clinical pathology), in order to control, to the extent possible, any extraneous variance due to differences in this attribute variable, in conjunction

Table 2

Aptitude Examination Items: Identification by Source

Kit of Factor-Referenced Cognitive Tests		
Examination/(Factor)	Part	Item
Hidden Figures Test (CF1) (Flexibility of closure)	1	12
Opposites Test (FA2) (Associational fluency)	2	5
Making Sentences Test (FE1) (Expressional fluency)	2	18
Things Categories Test (FI3) (Ideational fluency)	2	-
Locations Test (I2) (Inductive reasoning)	1	5
Figure Classification Test (I3) (Inductive reasoning)	1	7
Calendar Test (IP1) (Integrative processes)	1	9
First and Last Names Test (MA3) (Associative memory)	1	-
Number Comparison Test (P2) (Perceptual speed)	1	10
Necessary Arithmetic Operations Test (RG3) (General reasoning)	1	12
Nonsense Syllogisms Test (RL1) (Logical reasoning)	1	2
Inference Test (RL3) (Logical reasoning)	1	9

(table continues)

Examination/(Factor)	Part	Item
Deciphering Languages Test (RL4) (Logical reasoning)	1	4
Surface Development Test (VZ3) (Spatial visualization)	2	8
Making Groups Test (XU3) (Flexibility of use)	1	2

GRE General (Aptitude) Test

Section/(Item Type)	Part	Item
Analytical Ability (Analytical reasoning)	V	19
Analytical Ability (Logical reasoning)	V	24
Analytical Ability (Logical reasoning)	V	25
Verbal Ability (Analogies)	II	10
Verbal Ability (Sentence completion)	I	3

with the fact that potential subjects within the discipline of clinical pathology fulfilling the collective attribute variables of the matrix sampling strategy were available by means of personal and professional contacts to the investigator. The academic/occupational discipline attribute variable was likewise dichotomized into nonphysician and physician clinical pathology professions (i.e., medical technologists and pathologists, respectively). On the basis of informal pilot studies conducted prior to and in preparation for the present study, the maximum sample size of subjects feasible was determined to be four. The distribution of the sample of subjects in terms of the matrix sampling strategy attribute variables is presented in Table 3. Subjects satisfying the matrix sampling strategy attribute variables were identified and agreed to participate in the present study as volunteers. For purposes of identification, subjects were randomly assigned the arbitrary identification numbers of 101, 102, 103, and 104.

Procedure

The nonschedule standardized interview was individually administered to each of the four subjects for each of the 25 items in a series of sessions conducted during the summer of 1983. The sessions were scheduled at the convenience of the subjects, at approximately one week time intervals; the length of each session was at the discretion of each sub-

Table 3

Subjects: Distribution of Matrix Sampling Strategy Attribute Variables

Discipline			
Clinical Pathology			
Profession			
Medical Technologist or Pathologist			
Degree	Age	Sex	n
Baccalaureate	35-40	Male	1
	51-56	Female	1
Postbaccalaureate	35-40	Female	1
	51-56	Male	1

ject. The length of each session and the number of items administered per session were the determinants of the total number of sessions required of each subject. The number of sessions conducted with each subject varied from three to five, the length of the sessions varied from one to two hours, and the number of items administered per session varied from five to eight. All such sessions were conducted by the investigator.

At the outset of the first session with each subject, all relevant details concerning the present study were systematically and comprehensively reviewed, both as a means of orientation and as the means of securing the informed consent of each subject. Within the context of informed consent, subjects were advised:

- that the purpose of the study was to obtain data concerning the manner in which they responded to items of the type traditionally included on intelligence or academic aptitude examinations;
- that no inferences regarding their "intelligence" were capable of being derived, given the restricted sample of items to be administered and the lack of the investigator's formal "intelligence testing" training;
- that their anonymity would be maintained at all times during and subsequent to the sessions;
- that their responses to the nonschedule standardized interview would be tape-recorded, in order that transcripts of their responses, necessary for data analysis, could be prepared;
- that neither any potential risks nor benefits were anticipated to be experienced by them as a consequence of their participation as subjects in the study; and
- that they had the option to discontinue participation as subjects, without prejudice, at any time during the study.

During the orientation phase of the first session, a

copy of the nonschedule standardized interview was presented to and discussed with each subject. To sensitize, yet not bias subjects, in terms of the types of responses possible to the nonschedule standardized interview, for any aspects of the nonschedule standardized interview requiring clarification or elaboration, relevant illustrations were provided by the investigator within the context of clinical pathology. With respect to the procedure to be followed in conducting the nonschedule standardized interview, subjects were informed at this time:

- that each item, prefaced by the directions for that item, would be presented on a separate sheet of paper and would be posed in either a selected or constructed response format;
- that they would be provided with a pencil and that they were free to utilize the sheet of paper on which the item was presented as "scratch paper";
- that no significance was attributable to the sequence in which the items were presented, as the order of the items had been determined by means of a table of random numbers;
- that the items would be encountered only in the order in which they were presented and would be encountered on a "one-time-only" basis;
- that the items varied in terms of "difficulty", and consequently, the possibility existed that subjects might be unable to respond to each and every item;
- that a conscientious attempt to respond to each item was imperative, as even the manner in which subjects determined they were unable to respond to any item would provide data relevant to the purpose of the study;
- that of more importance than whether or not their response to the item was "correct" or "incorrect" was the specificity and comprehensiveness with which subjects detailed the manner in which they were responding to the item;
- that both "covert" activities (e.g., "I'm pausing because I'm not sure what this sentence means." as well as "overt" activities (e.g., "I'm drawing a diagram on the page to help me figure out what information I'm missing for this question.") were to be detailed when

responding to the item;
- that no time limits were imposed for any of the items, that the amount of time expended on any item was at their discretion, and should be the amount of time and/or effort they considered to constitute a conscientious attempt at reponding to the item;
- that any clarification or elaboration of their responses requested by the investigator was not to be misconstrued as an indication that their responses were, in any way, incorrect or inadequate;
- that certain of the inquiries of the nonschedule standardized interview might seem redundant or repetitive of other inquiries or of responses already provided by the subjects and that any redundancy or repetitiveness was not to be misconstrued as an indication that their responses were, in any, incorrect or inadequate;
- that no feedback information would be provided concerning whether their response to any item was "correct" or "incorrect", in order to reduce the possibility that the manner in which they responded to any of the subsequent items might inadvertently be influenced by such feedback; and
- that there were considered to be no "good" or "bad" responses to any portion of the nonschedule standardized interview and that they should not hesitate to be candid in their responses.

The tape-recordings of the nonschedule standardized interview with each subject for each item were subsequently transcribed verbatim and unedited by the investigator. For a one-hour session, approximately eight hours of time were required to completely transcribe the tape-recording of that session and to verify or "proofread" the resultant transcript against the tape-recording. Transcripts of the nonschedule standardized interview for each subject for each item were typed single-spaced with a pica element typewriter. Each transcript averaged five typewritten pages in length, with the number of pages varying from two to eight. The resultant transcripts (n = 100), constituting the data base for

the present study were subsequently content analyzed by the investigator.

Content Analysis

Content analysis was utilized as the means of reducing and analyzing the transcripts which constituted the data base for the present study, given the qualitative, rather than quantitative, type of data and the appropriateness of content analysis for qualitative data (Krippendorff, 1980; Patton, 1980; Newell and Simon, 1972). A representative and concise definition of content analysis is as follows:

> analysis of the manifest and latent content of a body of communicated material ... through a classification, tabulation, and evaluation of its key symbols and themes in order to ascertain its meaning and probable effect (Webster's Ninth New Collegiate Dictionary, 1983, p. 283).

The content analysis was conducted in an inductive manner, in that:

> ... the patterns, themes, and categories of analysis ... emerge[d] out of the data rather than being imposed on [the data] prior to data collection and analysis (Patton, 1980, p. 306).

That is, other than having presumed that content analysis of the transcripts of the thinking-aloud responses of the subjects would yield data relevant to the validity and reliability of the items and data relevant to assessing the internal and external validity of the results of the present study, no preconceived assumptions had been formulated concerning the specific types of information that would result from the content analysis.

Attempts to reduce the data into meaningful and manageable form eventually resulted in what were termed psychometric inferences, or inferences relevant to the validity and reliability of the items. Three general types of psychometric inferences were suggested by the data and were designated content/construct validity, internal consistency/discrimination, and alternate form/test-retest reliability. The definitions ascribed to the terms of the three types of psychometric inferences were analogous to the definitions of the terms within the context of psychometric theory. That is, content/construct validity encompassed aspects relevant to the sources of variance or determinants of "what" was measured by the item (e.g., aptitude(s), achievement) as well as presumably extraneous sources of variance, relative to the aptitude presumed to be measured by the item (e.g., ambiguity in the directions provided for the item, ambiguity in the item). Internal consistency/discrimination encompassed the sources of variance or determinants which served to differentiate between and among subjects with respect to item scores (e.g., aptitude(s), strategies, random errors of measurement[1]). Alternate form/test-retest

[1]Analogous to the denotation in psychometric theory, random errors of measurement in the present study reflected differences between "true" and "obtained" scores. Random errors of measurement were considered to exist when the item response selected or generated by a subject to an item did not parallel the thinking-aloud response of the subject to the item (i.e., the "right answer for the wrong reason", a false positive; the "wrong answer for the right reason", a false negative).

reliability encompassed aspects relevant to presumably parallel items within a so-called "factor-pure" examination and presumably parallel items which varied in response formats (e.g., parallels in content/construct validity, parallels in internal consistency/discrimination, parallels in "difficulty"[2]).

The psychometric inferences for each item were compared to the psychometric data available for each item to assess the extent to which the psychometric inferences supplemented the psychometric data. The psychometric data for each item were restricted to the operational definition of the aptitude purported to be measured by the item, as no other psychometric data (e.g., item analysis indices) were available or obtainable. Operational definitions for the items were considered psychometric data, although nonquantitative and descriptive data, in that operational definitions explicitly and implicitly reflect psychometric attributes. Furthermore, operational definitions are traditionally referenced to quantitative psychometric indices a priori and/or a posteriori to aptitude measure construction and calibration procedures and seemingly represent a direct extension of

[2]The "difficulty" of parallel items was not restricted to the denotation of the corresponding item analysis index in psychometric theory. For the purpose of the present study, the difficulty of an item referred to various aspects of the item which would tend to decrease the probability of a subject responding correctly to that item (e.g., inclusion of words not familiar to the subject, the "complexity" of a geometric figure constituting the basis of an item).

psychometric coefficients. Hence, considering the operational definitions as psychometric data was deemed justifiable.

For certain aspects of the items, presumably relevant to the validity and reliability of the items, no opportunity existed to derive psychometric inferences by means of the content analysis of the transcripts. Subjects had been presented with only the directions for the item and the item, on a single sheet of paper, during the individual administrations of the nonschedule standardized interview. Therefore, no opportunity existed to obtain the thinking-aloud responses of the subjects concerning whether the "practice" items, included in the source examination but not reproduced for subjects, were "beneficial"; whether having to "flip back and forth" between the directions for an item, located on an examination booklet cover sheet, and the item, located within an examination booklet, was distracting and/or cumbersome for subjects; among others. No explicit opportunity was available for assessing the directions provided for the scoring of items, which would presumably be relevant to the validity and reliability of the item scores. Aspects such as those delineated above were addressed by means of a rational analysis by the investigator and were included among the psychometric inferences derived from the content analysis of the transcripts of subjects' responses.

Attempts to assess the internal and external validity

of the results of the present study resulted in what were termed methodological inferences. Further content analysis of the transcripts and within-method and between-method triangulation were utilized to validate the three principal components of the present study: the subjects as the data sources, the nonschedule standardized interview as the means of data collection, and the investigator as the content analyst (Denzin, 1978; Lieberman, 1979; Krippendorff, 1980; Patton, 1980). The criteria to be utilized in assessing the internal and external validity of the data source, data collection, and data analysis components were suggested by the literature relevant to intelligence research, the literature relevant to qualitative research, and by the data. Within-method triangulation, relevant to internal validity, was assessed by means of the following criteria: whether the manner in which subjects anticipated responding to the item paralleled the manner in which subjects responded to the item and whether the responses of the subjects to the various aspects of the nonschedule standardized interview revealed interindividual differences in content and comprehensiveness. The first criterion was seemingly consistent with the task analytic premise which served as the basis for the content of the first portion of the nonschedule standardized interview; the second criterion was consistent with the interindividual differences premise intrinsic to intelligence theory research. Between-method triangulation, relevant to

the external validity, was assessed by means of the following criteria: whether the strategies utilized by subjects in responding to the item paralleled those described in the literature, for applicable items; whether the manner in which subjects responded to the item paralleled the task analysis described in the literature, for applicable items; and whether the psychometric inferences derived for the item paralleled those described in the literature, for applicable items. All three criteria were seemingly consistent with aspects documented in the literature. To the extent that the respective criteria delineated were consistent with the data, the internal and external validity were considered to be supported.

CHAPTER IV

RESULTS AND DISCUSSION

The results of the present study are at the level of the items, however, are most feasibly presented in terms of the factor purported to be measured by the item, as more than one item may constitute a measure of certain factors. The order of presentation of the factors is arbitrary. Preceding the results for each item, two types of descriptive information are provided. The first type of descriptive information is that of the operational definition of the factor purported to be measured by each item, as provided in the Manual and Bulletin for the two source examinations from which the items were selected (i.e., Kit of Factor-Referenced Cognitive Tests, Ekstrom et al., 1976b; Graduate Record Examination, Educational Testing Service, 1982; respectively). The second type of information is that of a brief description of the item(s) selected to constitute measure(s) of each factor, abbreviated as in Table 4. Such descriptive information is provided as a context within which to present the two types of results derived for each item (i.e., psychometric and methodological inferences).

As was previously discussed in the Content Analysis section of the Methodology chapter, three types of psycho-

Table 4

Aptitude Examination Items: Abbreviations

Kit of Factor-Referenced Cognitive Tests	
Examination/(Factor)	Abbreviation
Hidden Figures Test (CF1) (Flexibility of closure)	KIT/CF1/1/12
Opposites Test (FA2) (Associational fluency)	KIT/FA2/2/5
Making Sentences Test (FE1) (Expressional fluency)	KIT/FE1/2/18
Things Categories Test (FI3) (Ideational fluency)	KIT/FI3/2/-
Locations Test (I2) (Inductive reasoning)	KIT/I2/1/5
Figure Classification Test (I3) (Inductive reasoning)	KIT/I3/1/7
Calendar Test (IP1) (Integrative processes)	KIT/IP1/1/9
First and Last Names Test (MA3) (Associative memory)	KIT/MA3/1/-
Number Comparison Test (P2) (Perceptual speed)	KIT/P2/1/10
Necessary Arithmetic Operations Test (RG3) (General reasoning)	KIT/RG3/1/12
Nonsense Syllogisms Test (RL1) (Logical reasoning)	KIT/RL1/1/2
Inference Test (RL3) (Logical reasoning)	KIT/RL3/1/9

(table continues)

Examination/(Factor)	Abbreviation
Deciphering Languages Test (RL4) (Logical reasoning)	KIT/RL4/1/4
Surface Development Test (VZ3) (Spatial visualization)	KIT/VZ3/2/8
Making Groups Test (XU3) (Flexibility of use)	KIT/XU3/1/2

GRE General (Aptitude) Test

Section/(Item Type)	Abbreviation
Analytical Ability (Analytical reasoning)	GRE/AAR/V/19
Analytical Ability (Logical reasoning)	GRE/ALR/V/24
Analytical Ability (Logical reasoning)	GRE/ALR/V/25
Verbal Ability (Analogies)	GRE/VAN/II/10
Verbal Ability (Sentence completion)	GRE/VSC/I/3

metric inferences were derived for each item, designated as content/construct validity, internal consistency/discrimination, and alternate form/test-retest reliability. Content/construct validity was defined to encompass aspects relevant to the sources of variance or determinants of "what" was measured by the item (e.g., aptitude(s), achievement) as well as presumably extraneous sources of variance, relative to the aptitude purported to be measured by the item (e.g., ambiguity in the directions provided for the item, ambiguity in the item). Internal consistency/discrimination was defined to encompass those sources of variance or those determinants which served to differentiate between and among subjects with respect to item scores (i.e., aptitude(s); strategies; random errors of measurement, as was defined in Note 1 in the Content Analysis section of the Methodology chapter). Alternate form/test-retest reliability was defined to encompass aspects relevant to presumably parallel items within a so-called "factor-pure" examination and presumably parallel items which varied in response formats (e.g., parallels in content/construct validity; parallels in internal consistency/discrimination; parallels in difficulty, as was defined in Note 2 in the Content Analysis section of the Methodology chapter). Providing illustrative excerpts from the transcripts of subjects' responses to either selectively or comprehensively support the psychometric inferences for each item is precluded by virtue of the voluminous text that such

documentation would require. Consequently, the psychometric inferences for each item are summarized in tabular form (see respective odd numbered tables entitled Summary of Psychometric Inferences).

As was likewise previously discussed in the Content Analysis section of the Methodology chapter, two types of methodological inferences were derived for each item, designated as within-method triangulation and between-method triangulation. Within-method and between-method triangulation were utilized to assess the internal and external validity, respectively, of the three principal components of the present study: the subjects as the data sources, the nonschedule standardized interview as the means of data collection, and the investigator as the content analyst. The criteria utilized to assess the internal validity of the data source, data collection, and data analysis components included whether the manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item and whether the responses of the subjects to the various aspects of the nonschedule standardized interview revealed interindividual differences in content and comprehensiveness. The criteria utilized to assess the external validity of the data source, data collection, and data analysis components included whether the strategies utilized by subjects in responding to the item paralleled those described in the literature, for applicable items; whether the

manner in which subjects responded to the item paralleled the task analysis described in the literature, for applicable items; and whether the psychometric inferences derived for the item paralleled those described in the literature, for applicable items. Providing illustrative excerpts from the transcripts of subjects' responses to either selectively or comprehensively support the methodological inferences for each item is precluded by virtue of the voluminous text that such documentation would require. Consequently, the methodological inferences for each item are summarized in tabular form (see respective even numbered tables entitled Summary of Methodological Inferences).

Verbal Ability (Sentence Completion)

Operational Definition of the Items

Provided in the GRE 1982-83 Information Bulletin (Educational Testing Service, 1982) was the following description and discussion of sentence completion items, presumed to be equivalent to an operational definition.

> The purpose of the sentence completion questions is to measure the ability to recognize words ... that both logically and stylistically complete the meaning of a sentence. In deciding which ... words can best be substituted for blank spaces in a sentence, one must analyze the relationships among the component parts of the incomplete sentence. One must consider each [word] and decide which completes the sentence in such a way that the sentence has a logically satisfying meaning and can be read as a stylistically integrated whole. Sentence completion questions provide a context within which to analyze the function of words as they relate to and combine with one another to form a meaningful unit of discourse (p. 11).

No further information relevant to the operational definition or other psychometric attributes of verbal ability, in the form of sentence completion items, was provided.

Description of the Items

GRE/VSC/I/3.

The stem of this item consisted of a sentence from which two words had been omitted and had been replaced by blank spaces, represented by a series of hyphens (i.e., ---). Five nonexhaustive options were provided for the item; each option consisted of a pair of words. Subjects were to select the pair of words which, when substituted into the blank spaces in the stem of the item, was most consistent with the meaning of the sentence. This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

GRE/VSC/I/3a.

This item was modified from the former item by the investigator so as to be posed in a constructed/unrestrictive response format, the former item having been posed in a selected/nonexhaustive response format. The stem of this item was identical to that of the former item, except that the blank spaces had been represented by lines (i.e., _____). Subjects were to generate words which, when substituted into each of the blank spaces in the stem of the item, were most consistent with the meaning of the sentence. The directions

for this item were modified from those of the former item only so as to be consistent with the constructed response format of this item.

Inferences Relevant to the Items

Psychometric inferences relevant to item GRE/VSC/I/3 are summarized in Table 5; methodological inferences relevant to item GRE/VSC/I/3 are summarized in Table 6. Psychometric inferences relevant to item GRE/VSC/I/3a are summarized in Table 7; methodological inferences relevant to item GRE/VSC/I/3a are summarized in Table 8.

Verbal Ability (Analogies)

Operational Definition of Item GRE/VAN/II/10

Provided in the GRE 1982-83 Information Bulletin (Educational Testing Service, 1982) was the following description and discussion of verbal analogy items, presumed to be equivalent to an operational definition.

> Analogy questions test the ability to recognize relationships among words and the concepts they represent and to recognize when these relationships are parallel. The [questions] require one to formulate and then to analyze the relationships linking ... pairs of words and to recognize which ... [relationships are] most nearly analogous (p. 9).

No further information relevant to the operational definition or other psychometric attributes of verbal ability, in the form of analogy items, was provided.

Description of Item GRE/VAN/II/10

The stem of the item consisted of a pair of words pre-

Table 5

GRE/VSC/I/3: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - reading comprehension; - vocabulary; - efficiency of responding (i.e., "speededness"); - "recall" rather than "recognition" or both "recall" and "recognition" (i.e., of appropriate words for the blank spaces in the stem of the item). The operational definition for the item stated that the item measured the ability to recognize words that "... logically ... complete the meaning of a sentence". Whether or not a "logical" criterion was utilized in selecting words for the blank spaces is indeterminate, however, a "semantic" criterion was definitely utilized. The directions for the item were not sufficiently explicit that the two words to be selected for the blank spaces in the stem of the item must be contained within one and only one option, rather than one word from one option and one word from another option. The directions for the item were not sufficiently explicit that the first word of each option corresponded to the first blank space and that the second word of each option corresponded to the second blank space in the stem of the item.
Internal Consistency/Discrimination
The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 2, incorrect responses = 2). Aptitudes constituted a source of variance in item scores,

(table continues)

Internal Consistency/Discrimination

as interindividual differences in item scores were attributable to the aptitudes previously delineated.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty:

- dependent upon the style in which the sentence was written (e.g., "concrete" versus "abstract");
- dependent upon whether the words contained in the stem and/or options of the item were familiar or unfamiliar.
- dependent upon the response format of the item (i.e., selected versus constructed response format, specifically, "recognition" versus "recall").

(See also the corresponding section for item GRE/VSC/I/3a.)

Table 6

GRE/VSC/I/3: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - The strategies utilized by subjects in responding to the item paralleled those described in the literature (Bloom and Broder, 1950; Olshavsky, 1976-1977; Kavale and Schreiner, 1979; Fareed, 1971; Educational Testing Service, 1982). - (A task analysis relevant to this item was not identified.)

Table 7

GRE/VSC/I/3a: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- reading comprehension;
- vocabulary;
- efficiency of responding (i.e., "speededness");
- "recall" rather than "recognition" (i.e., of appropriate words for the blank spaces in the stem of the item, although consistent with the constructed response format);
- verbal closure (i.e., "[t]he ability to solve problems requiring the identification of visually presented words when some of the letters are missing, scrambled, or embedded among other letters"; Ekstrom et al., 1976b, p. 33), if considered at the level of sentences and words, rather than words and letters;
- expressional fluency (see operational definition for item KIT/FE1/2/18).

The operational definition for the item stated that the item measured the ability to recognize (i.e., recall for this item) words that "... stylistically complete the meaning of a sentence". Whether or not a "stylistic" criterion was utilized in generating words for the blank spaces is indeterminate, however, a "stylistic" criterion was applied to a word contained in the stem of the item.

Internal Consistency/Discrimination

The item responses of the subjects did not reveal interindividual differences in terms of item scores (i.e., correct responses = 4, incorrect responses = 0), with the "correctness" of item responses subjectively assessed by the investigator.

(table continues)

Internal Consistency/Discrimination

Aptitudes did not constitute a source of variance in item scores, although interindividual differences in the aptitudes previously delineated were exhibited.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel item would vary in difficulty:

- dependent upon the style in which the sentence was written (e.g., "concrete" versus "abstract");
- dependent upon whether the words contained in the stem of the item were familiar or unfamiliar;
- dependent upon the response format of the item (i.e., selected versus constructed response format, specifically, "recognition" versus "recall").

In conjunction with item GRE/VSC/I/3, the two forms of this item were not parallel in terms of content/construct validity (see respective sections for both items). With respect to internal consistency/discrimination, interindividual differences and intra-individual consistencies were not parallel between the two forms of the item (see respective sections for both items). In terms of alternate form/test-retest reliability, the sources of difficulty between the two forms of the item were parallel (see respective sections for both items).

Table 8

GRE/VSC/I/3a: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered to have been supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered to have been supported by the following criterion: - The strategies utilized by subjects in responding to the item paralleled those described in the literature (Bloom and Broder, 1950; Olshavsky, 1976-1977; Kavale and Schreiner, 1979; Fareed, 1971; Educational Testing Service, 1982). - (A task analysis relevant to this item was not identified.)

sented in analogical notation (i.e., A:B::) which were synonyms. Five nonexhaustive options were provided for the item, each option likewise consisted of a pair of words presented in analogical notation (i.e., C:D). Subjects were to select the pair of words which "... expresse[d] a relationship similar to that expressed in the original pair [of words constituting the stem of the item]" (Educational Testing Service, 1982, p. 28). The term "verbal analogy" did not appear in the directions for the item. That an option other than the correct, keyed option could likewise be justified as a correct response to the item had been suggested both by inspection and by the responses of subjects to the item in the pilot studies conducted prior to and in preparation for the present study; this item was selected from among the items contained within the source examination for that reason.

Inferences Relevant to Item GRE/VAN/II/10

Psychometric inferences relevant to this item are summarized in Table 9. Methodological inferences relevant to this item are summarized in Table 10.

Associational Fluency

Operational Definition of Item KIT/FA2/2/5

Provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition of associational fluency: "The ability

Table 9

GRE/VAN/II/10: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - vocabulary; - familiarity with the type of item (i.e., "achievement" versus "aptitude"). The directions for the item did not define or otherwise explain the analogical notation utilized in the stem and options of the item (i.e., A:B::, C:D, respectively). Given the clarity of the directions relevant to the task posed by the item, inclusion of the analogical notation in the item was superfluous and distracting to subjects unfamiliar with the notation. That an option other than the correct, keyed option could likewise be justified as a correct response to the item was further supported in that the explanations provided by subjects as justification of their item responses for both of the two options were "valid". "Some approaches that may be helpful in answering analogy questions" (p. 9) were provided in the description/discussion section of the Bulletin, however, none of these approaches were reiterated in the directions for the item.
Internal Consistency/Discrimination
The item responses of the subjects revealed interindividual differences in terms of item scores, based on the correct, keyed option for the item (i.e., correct responses = 1, incorrect responses = 3). Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attrib-

(table continues)

Internal Consistency/Discrimination

utable to the aptitudes previously delineated.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement constituted a source of variance in item scores, as the three incorrect item responses were analogous to the "wrong answer for the right reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the type of relationship embodied in the pairs of words constituting the stem and options of the item;
- whether the words contained in the stem and options of the item were familiar or unfamiliar.

Table 10

GRE/VAN/II/10: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The strategies utilized by subjects in responding to the item paralleled those described in the literature (Pellegrino and Glaser, 1979; Sternberg, 1974; Educational Testing Service, 1982). - The responses of the subjects to the item paralleled the task analysis identified in the literature (Sternberg, 1974).

to produce rapidly words which share a given area of meaning or some other semantic property" (p. 41). No further information relevant to the operational definition or other psychometric attributes of associational fluency was provided.

Description of Item KIT/FA2/2/5

This item presented an adjective as the stem of the item. Subjects were to list a maximum of six antonyms for the word constituting the stem of the item on the six blank lines provided. No restrictions were specified relative to permissable responses to the item. This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

Inferences Relevant to Item KIT/FA2/2/5

Psychometric inferences relevant to this item are summarized in Table 11. Methodological inferences relevant to this item are summarized in Table 12.

Expressional Fluency

Operational Definition of Item KIT/FE1/2/18

Provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition of expressional fluency: "The ability to think rapidly of word groups or phrases" (p. 51). No further information relevant to the operational definition

Table 11

KIT/FA2/2/5: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - vocabulary; - distractibility (i.e., ability to attend to antonyms rather than synonyms); - ease of retrieval of words from memory and/or hierarchical clustering/chunking of semantic memory - compulsivity (i.e., striving to list six antonyms merely because six blank lines were provided). The directions for the item did not emphasize (e.g., capital letters, underlining) that antonyms, rather than synonyms, were required as responses. Neither did the directions for the item explicitly state whether only single words were acceptable as responses or whether word combinations (e.g., two-word phrases) were acceptable as responses. The directions did not specify whether or not the antonyms generated as responses had to be spelled correctly and/or had to conform to the part of speech (e.g., noun, adjective) represented by the word constituting the stem of the item. The directions for scoring the item did not specify guidelines or criteria for assessing the "correctness" of responses to the item (e.g., spelling, semantics, parts of speech).
Internal Consistency/Discrimination
The item responses of the subjects revealed negligible interindividual differences in terms of item scores, based only on the number of words listed (i.e., number of antonyms listed = 4, 6, 6, 4). Aptitudes did not constitute a source of variance in item

(table continues)

Internal Consistency/Discrimination

scores, although interindividual differences in the aptitudes previously delineated were exhibited.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- whether the word constituting the stem of the item was familiar or unfamiliar;
- the response format of the item (i.e., selected versus constructed response format, specifically, "recognition" versus "recall").

Item scores as well as examination scores would vary within and between both investigators and studies dependent upon the criteria utilized in assessing the "correctness" of item responses.

Table 12

KIT/FA2/2/5: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed interindividual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - The strategies utilized by subjects in responding to the item paralleled those described in the literature (Bloom and Broder, 1950; Olshavsky, 1976-1977; Kavale and Schreiner, 1979; Fareed, 1971; Educational Testing Service, 1982). - (A task analysis relevant to this item was not identified.)

or other psychometric attributes of expressional fluency was provided.

Description of Item KIT/FE1/2/18

The stem of this item consisted of six blank lines preceded by either letters of asterisks; the sixth blank line was followed by a period. Subjects were to write a sentence by placing a word in each of the blank lines. For the three blank lines preceded by a letter, the word placed in each blank line was required to begin with that letter; for the three blank lines preceded by an asterisk, the word placed in the blank lines was permitted to begin with any letter. Restrictions were imposed on the words to be placed in the blank lines (e.g., abbreviations were not acceptable, contractions were acceptable). This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

Inferences Relevant to Item KIT/FE1/2/18

Psychometric inferences relevant to this item are summarized in Table 13. Methodological inferences relevant to this item are summarized in Table 14.

Ideational Fluency

Operational Definition of Item KIT/FI3/2/-

Provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) was the following

Table 13

KIT/FE1/2/18: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- vocabulary;
- grammar and/or sentence structure rule knowledge;
- capacity of memory (i.e., for restrictions specified in the directions for the item);
- reading comprehension (i.e., of the directions for the item);
- "innovativeness" and/or "improvisation" in written expression.

The directions for the item were lengthy, contained numerous specifications to be considered in constructing acceptable responses, and seemed to lack continuity and/or were "disjointed". Consequently, numerous readings of the directions were required prior to responding to the item.

The directions for the item were ambiguous with respect to certain of the specifications to be considered in constructing acceptable responses to the item. Although the directions included the word "sentence" on all relevant occasions and although the final blank line was followed by a period, not explicitly stated was whether "questions" as well as sentences constituted acceptable responses to the item. The term "proper names" was included in the directions for the item, however, the term was not defined explicitly or implicitly by means of the examples of proper names included in the directions. Consequently, inquiries from the subjects could not be addressed merely from reading the directions (e.g., whether a day of the week was a proper name).

The directions for scoring the item did not specify guidelines or criteria for assessing the "correctness" of item responses, in terms of the above ambiguities. Furthermore, the directions for scoring delineated additional criteria to be utilized in assessing the "correctness" of item responses which had not, however, been communicated to sub-

(table continues)

Content/Construct Validity

jects in the directions for the item.

In actual administration of the examination, rather than merely one item, the directions for the examination would be presented on the equivalent of an examination booklet cover sheet. Given the length of and the specifications in the directions, subjects would be required to "flip back and forth" between the cover sheet and items in order to refer to the directions. The extent to which referring to the examination booklet cover sheet would be distracting and/or time-consuming is indeterminate.

Internal Consistency/Discrimination

The item responses of the subjects revealed no interindividual differences in terms of item scores, when no criteria other than those explicitly provided in the directions for scoring the item were utilized (i.e., acceptable "sentences" or "questions" = 4; unacceptable sentences or questions = 0).

Aptitudes did not constitute a source of variance in item scores, and no appreciable interindividual differences in the aptitudes previously delineated were exhibited.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the number of blank lines prefaced with letters rather than with asterisks.

Item scores as well as examination scores would vary within

(table continues)

and between both investigators and studies dependent upon the criteria utilized in assessing the "correctness" of item responses.

Table 14

KIT/FE1/2/18: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The strategies utilized by subjects in responding to the item paralleled those described in the literature (Carroll, 1976; Ekstrom et al., 1976b). - (A task analysis relevant to this item was not identified in the literature.) - A review of other studies having utilized various combinations of the so-called "marker tests" of expressional fluency contained within the source examination resulted in the conclusion that "... the expressional fluency factor appears to have little support" (Ekstrom, French, and Harman, 1979, p. 16), perhaps attributable to the multivariate sources of variance in conjunction with the confounding source of variance of the directions for the item.

operational definition for ideational fluency: "The facility to write a number of ideas about a given topic or exemplars of a given class of objects" (p. 67). No further information relevant to the operational definition or other psychometric attributes of ideational fluency was provided.

Description of Item KIT/FI3/2/-

The stem of the item specified a concept (e.g., a shape) for which subjects were to list as many "things" as possible of that shape. Consistent with the constructed response format of the item, thirty-six blank lines were provided on which subjects were to list or otherwise describe the "things" in one or more words. No restrictions were specified relative to permissible item responses. This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

Inferences Relevant to Item KIT/FI3/2/-

Psychometric inferences relevant to this item are summarized in Table 15. Methodological inferences relevant to this item are summarized in Table 16.

General Reasoning

Operational Definition of the Items

Provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition for general reasoning: "The ability

Table 15

KIT/FI3/2/-: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition of the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- "imagination";
- concentrating ability;
- vocabulary;
- concept differentiation and/or acculturation;
- ease of retrieval of words from memory and/or hierarchical clustering/chunking of semantic as well as figural memory;
- associational fluency (see operational definition for item KIT/FA2/2/5);
- efficiency of responding (i.e., "speededness").

The directions for the item included the phrase "... things that are [specified shape] or that are [specified shape] more often than any other shape". The phrase was confusing in that one subject presumed he/she was to list "things" which changed shapes by changing physical states (e.g., ice cubes, solid "squares", melt to form water, liquid "round" puddles).

The directions for the item provided neither guidelines nor criteria concerning the extent to which a generic concept did or did not preclude listing specific examples of that concept (e.g., did listing "ball" preclude listing basketball, baseball, beach ball?). Subjects were required to interpret "what the directions probably meant".

The directions provided for scoring the item did not specify explicit or implicit criteria for assessing the "correctness" of item responses (e.g., was an oval "thing" equivalent to a round "thing"?).

(table continues)

Internal Consistency/Discrimination

The item responses of the subjects revealed marginal interindividual differences in terms of item scores, when no criteria other than the number of "things" listed were utilized (i.e., number of "things" listed = 17, 16, 16, 11).

Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement constituted a source of variance in item scores, as the item responses of subjects who presumed that listing a generic concept precluded listing specific examples of the concept were analogous to the "wrong answer (omitted, and hence no credit) for the right reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the familiarity or unfamiliarity of the concept specified in the stem of the item;
- the number of potential responses in the domain of responses for a given concept.

Item scores as well as examination scores would vary within and between subjects dependent upon whether subjects presumed that listing a generic concept precluded listing examples of that concept.

Item scores as well as examination scores would vary within and between both investigators and studies dependent upon the criteria utilized in assessing the "correctness" of item responses.

Table 16

KIT/FI3/2/-: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The strategies utilized by subjects in responding to the item paralleled those described in the literature (Frederiksen, 1969; Bower and Hilgard, 1981; Carroll, 1976). - The responses of the subjects to the item paralleled the task analysis identified in the literature (Bower and Hilgard, 1981; Frederiksen, 1969). - A review of other studies having utilized various combinations of so-called "marker tests" of ideational fluency resulted in the conclusion that "[t]here appears to be a good deal of confusion still surrounding this factor. .... [T]he more restrictive the stimulus, the greater the loading on associational fluency instead of ideational fluency" (Ekstrom et al., 1979, p. 18), perhaps accounting for the delineation of associational fluency as a source of variance for this item.

to select and organize relevant information for the solution of a problem" (p. 133). No further information relevant to the operational definition or other psychometric attributes of general reasoning was provided.

Description of the Items

KIT/RG3/1/12.

The stem of the item consisted of the particulars of an arithemtic/algebraic "story" or "word" problem. The four nonexhaustive options provided consisted of pairs of arithmetic operations (e.g., addition and subtraction) which represented possible means for solving the problem posed in the stem of the item. This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

KIT/RG3/1/12a.

This item was modified from the former item by the investigator so as to be posed in a constructed/unrestrictive response format, the former item having been posed in a selected/nonexhaustive response format. The stem of this item was identical to that of the former item. Subjects were to calculate the numerical solution to the problem and write the resultant solution on the blank line provided. The directions for this item were modified from those of the former item only so as to be consistent with the constructed response format of this item.

KIT/RG3/1/12b.

This item was modified from the former item by the investigator so as to be posed in a selected/exhaustive response format, the former item having been posed in a selected/nonexhaustive response format. The stem of this item was identical to that of the former item. Four of the five options provided for this item consisted of numerical solutions to the problem posed in the stem of the item; the fifth option provided consisted of a "none of the above" response. Subjects were to select the option which corresponded to the numerical solution for the problem posed in the stem of the item. The directions for this item were modified from those of the former item only so as to be consistent with the selected response format of this item.

Inferences Relevant to the Items

Psychometric inferences relevant to item KIT/RG3/1/12 are summarized in Table 17; methodological inferences relevant to item KIT/RG3/1/12 are summarized in Table 18. Psychometric inferences relevant to item KIT/RG3/1/12a are summarized in Table 19; methodological inferences relevant to item KIT/RG3/1/12a are summarized in Table 20. Psychometric inferences relevant to item KIT/RG3/1/12b are summarized in Table 21; methodological inferences relevant to item KIT/RG3/1/12b are summarized in Table 22.

Table 17

KIT/RG3/1/12: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition of the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- "recall" (i.e., rote application of arithmetic solution to the problem posed in the stem of the item);
- "reasoning" (i.e., formulation of algebraic solution to the problem posed in the stem of the item);
- efficiency of responding (i.e., "speededness");
- inductive reasoning (see operational definition for item KIT/I2/1/5);
- logical reasoning (see operational definition for item KIT/RL1/1/2).

The directions for the item specified that "[w]hen two [arithmetic] operations are given, they are always given in the order in which they should be performed". However, correctly solving the problem posed in the stem of the item was possible by employing the arithmetic operations in the reverse order of that given in the correct, keyed option. The specification in the directions for the item concerning the order of the arithmetic operations was distracting to the one subject employing the arithmetic operations in the reverse order of the order specified in the option.

The correct, keyed option for the item was not comprehensive and hence not entirely accurate. The arithmetic operations contained in the correct, keyed option omitted one operation necessary for the solution to the problem posed in the stem of the item (i.e., multiplication, to convert proportion to per cent).

The stem of the item phrased the essence of the problem by means of "What was the per cent reduction?", without explicitly explaining what was meant by the term. Subjects were thus required to be familiar with the term and to be further aware that the stem of the item was to be read as the "per cent reduction [in the price of an item]".

(table continues)

Content/Construct Validity

The directions provided for scoring in the source examination specified that examination scores would be "corrected for guessing" (i.e., scores equal to the number of items marked correctly minus a fraction of the number of items marked incorrectly). However, the rationale for such a scoring procedure was not provided in the description/discussion for the examination in the Manual. Furthermore, nowhere in the directions for scoring the examination was the fraction to be utilized in the "correction for guessing" formula specified.

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 3, incorrect responses = 1).

Aptitudes did not constitute a source of variance in item scores, although interindividual differences in the aptitudes previously delineated were exhibited.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement constituted a source of variance in item scores, as the incorrect response was analogous to the "wrong answer for the right reason", and one correct response was analogous to the "right answer for the wrong reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- whether the solution to the problem posed in the stem of the item entailed "recall" or "reasoning".

(table continues)

Alternate Form/Test-Retest Reliability

Item scores as well as examination scores would vary within and between both investigators and studies dependent upon the fraction utilized in the "correction for guessing" formula.

A parallel form of this item which might eliminate or reduce the possibility of subjects selecting the "right answer for the wrong reason" would be if the options provided consisted of the arithmetic/algebraic equations for possible solutions to the problem posed in the stem of the item (e.g., per cent reduction = [40.00 - 29.99] ÷ 40.00 x 100).

(See also the corresponding section for item KIT/RG3/1/12b.)

Table 18

KIT/RG3/1/12: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.

Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - (Strategies utilized by subjects in responding to the item were not identified in the literature.) - (A task analysis relevant to the item was not identified in the literature.) - No consensus seemingly exists concerning the extent to which general reasoning is or is not exclusive of other types of reasoning (e.g., logical, inductive) and/or arithmetic/numerical facility (French, 1957; Green et al., 1953; Carroll, 1976; Ekstrom et al., 1976b, 1979).

Table 19

KIT/RG3/1/12a: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - "recall" (i.e., rote application of arithmetic solution to the problem posed in the stem of the item); - "reasoning" (i.e., formulation of algebraic solution to the problem posed in the stem of the item); - efficiency of responding (i.e., "speededness"); - inductive reasoning (see operational definition for item KIT/I2/1/5); - logical reasoning (see operational definition for item KIT/RL1/1/2); - arithmetic/numerical facility, more so if subjects performed the required calculations "longhand" than with a calculator. The stem of the item phrased the essence of the problem by means of "What was the per cent reduction?", without explicitly explaining what was meant by the term. Subjects were thus required to be familiar with the term and to be further aware that the stem of the item was to be read as the "per cent reduction [in the price of an item]".
Internal Consistency/Discrimination
The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 2, incorrect responses = 2). Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated. Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies

(table continues)

Internal Consistency/Discrimination

were exhibited.

Random errors of measurement constituted a source of variance in item scores, as one incorrect response was analogous to the "wrong answer for the right reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- whether the solution to the problem posed in the stem of the item entailed "recall" or "reasoning".

(See also the corresponding section for item KIT/RG3/1/12b).

Table 20

KIT/RG3/1/12a: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - (Strategies utilized by subjects in responding to the item were not identified in the literature.) - (A task analysis relevant to the item was not identified in the literature.) - No consensus seemingly exists concerning the extent to which general reasoning is or is not exclusive of other types of reasoning (e.g., logical, inductive) and/or arithmetic/numerical facility (French, 1957; Green et al., 1953; Carroll, 1976; Ekstrom et al., 1976b, 1979).

Table 21

KIT/RG3/1/12b: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- "recall" (i.e., rote application of arithmetic solution to the problem posed in the stem of the item);
- "reasoning" (i.e., formulation of algebraic solution to the problem posed in the stem of the item);
- efficiency of responding (i.e., "speededness");
- inductive reasoning (see operational definition for item KIT/I2/1/5);
- logical reasoning (see operational definition for item KIT/RL1/1/2);
- arithmetic/numerical facility, more so if subjects performed the required calculations "longhand" than with a calculator.

The stem of the item phrased the essence of the problem by means of "What was the per cent reduction?", without explicitly explaining what was meant by the term. Subjects were thus required to be familiar with the term and to be further aware that the stem of the item was to be read as the "per cent reduction [in the price of an item]".

Internal Consistency/Discrimination

The item responses of subjects revealed no interindividual differences in terms of item scores (i.e., correct responses = 4, incorrect responses = 0).

Aptitudes did not constitute a source of variance in item scores, although interindividual differences in the aptitudes previously delineated were exhibited.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies

(table continues)

Internal Consistency/Discrimination

were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, although one correct response was analogous to the "right answer for the wrong reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- whether the solution to the problem posed in the stem of the item entailed "recall" or "reasoning".

In conjunction with items KIT/RG3/1/12 and KIT/RG3/1/12a, the three forms of this item were parallel in terms of content/construct validity, with the exception of arithmetic/numerical facility, not constituting a source of variance in the former item (see respective sections for all three items). With respect to internal consistency/discrimination, interindividual differences and intra-individual consistencies were not parallel across the three forms of the item (see respective sections for all three items). In terms of alternate form/test-retest reliability, the sources of difficulty among the three forms of the item were parallel (see respective sections for all three items).

Table 22

KIT/RG3/1/12b: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - (Strategies utilized by subjects in responding to the item were not identified in the literature.) - (A task analysis relevant to the item was not identified in the literature.) - No consensus seemingly exists concerning the extent to which general reasoning is or is not exclusive of other types of reasoning (e.g., logical, inductive) and/or arithmetic/numerical facility (French, 1957; Green et al., 1953; Carroll, 1976; Ekstrom et al., 1976b, 1979).

Logical Reasoning (GRE)

Operational Definition of the Items

Provided in the GRE 1982-83 Information Bulletin (Educational Testing Service, 1982) was the following description and discussion of logical reasoning, presumed to be equivalent to an operational definition.

> Logical reasoning questions test the ability to understand, analyze, and evaluate arguments. Some of the abilities tested by specific questions include recognizing the point of an argument, recognizing assumptions on which an argument is based, drawing conclusions from given premises, inferring material missing from given passages, applying principles governing one argument to another, identifying methods of argument, evaluating arguments and counterarguments, and analyzing evidence (p. 22).

No further information relevant to the operational definition or other psychometric attributes of logical reasoning was provided.

Description of the Items

GRE/ALR/V/24.

This item was based on what was termed an "argument", with the argument consisting of a conjunctive sentence of approximately 30 words in length. The five nonexhaustive options provided for this item likewise consisted of arguments, similar in length and construction to the argument constituting the basis for the item. Subjects were to select, from among the arguments provided as options, the argument which was most similar, in terms of "logical features", to the argument serving as the basis of this item. This

item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

GRE/ALR/V/25.

This item was likewise based on what was termed an argument consisting of a paragraph of approximately 80 words in length. Each of the five nonexhaustive options provided for this item consisted of a statement citing an instance or set of circumstances related to the content of the argument. Subjects were to determine which of the statements in the options would tend to weaken the argument. This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

Inferences Relevant to the Items

Psychometric inferences relevant to item GRE/ALR/V/24 are summarized in Table 23; methodological inferences relevant to item GRE/ALR/V/24 are summarized in Table 24. Psychometric inferences relevant to item GRE/ALR/V/25 are summarized in Table 25; methodological inferences relevant to item GRE/ALR/V/25 are summarized in Table 26.

Analytical Reasoning

Operational Definition of Item GRE/AAR/V/19

Provided in the GRE 1982-83 Information Bulletin (Educational Testing Service, 1982) was the following descrip-

Table 23

GRE/ALR/V/24: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for this item did not sufficiently acknowlege the following sources of variance as determinants in responding to the item: - general reasoning (see operational definition for item KIT/RG3/1/12); - inductive reasoning (see operational definition for item KIT/I2/1/5); - reading comprehension; - vocabulary; - capacity of memory (i.e., to retain the details contained within the arguments in the stem and options of the item); - familiarity with the type of item (i.e., "knowing what to look for" as well as being aware that the meanings of certain unfamiliar words/terms were irrelevant to responding to the item); - terminology and/or concepts of formal logic (i.e., "argument", "logical features"); - efficiency of responding (i.e., "speededness"). The length of the arguments constituting the basis of the item and constituting the options for the item necessitated numerous readings of the arguments prior to and while responding to the item. The directions for the item suggested a potentially advantageous strategy for responding to the item (i.e., drawing a "rough" diagram). Such a strategy, however, was not enumerated in the description/discussion for such items in the Bulletin. Other strategies had been enumerated in the same description/discussion, however, these other strategies were not reiterated in the directions for the item. No rationale for the selective listing of the one strategy in the directions for the item, at the exclusion of the other strategies, was provided.

(table continues)

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 3, incorrect responses = 1).

Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated.

Strategies constituted a source of variance in item scores, as interindividual differences in item scores were attributable to strategies utilized.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the complexity of the "logical features" embodied in the arguments;
- knowledge of the words contained in the arguments in the options and the argument serving as the basis of the item;
- prior exposure to and/or familiarity with the type of item.

(See also the corresponding section for item GRE/ALR/V/25.)

Table 24

GRE/ALR/V/24: Summary of Methodological Inferences

Within-Method Triangulation

The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria:

- The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item.
- The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.

Between-Method Triangulation

The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion:

- (Strategies utilized by subjects in responding to the item were not identified in the literature.)
- (A task analysis relevant to the item was not identified in the literature.)
- A review of other studies having utilized various so-called "marker tests" of logical reasoning resulted in the conclusion that few of such studies:

  > ... yielded a clear syllogistic or logical reasoning factor. [Certain examinations] tended to load on factors which also included induction tests ... [and certain other tests] tended to load on factors with vocabulary and/or general reasoning tests. This suggests that [logical reasoning] tests do not function similarly [across all subjects and administrations] ... (Ekstrom et al., 1979, p. 36).

Table 25

GRE/ALR/V/25: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- general reasoning (see operational definition for item KIT/RG3/1/12);
- reading comprehension;
- capacity of memory (i.e., to retain the details contained within the argument constituting the basis of the item);
- familiarity with the type of item (i.e., "knowing what to look for", within the contexts of reading or verbal comprehension as well as logical reasoning);
- terminology of formal logic (i.e., "argument");
- efficiency of responding (i.e., "speededness").

The length of the argument constituting the basis of the item necessitated numerous readings of the argument and options prior to and while responding to the item.

The directions for the item did not explicitly or implicitly state whether or not subjects were to assume any information beyond that presented in the argument. Whether subjects presumed that evolutionary stages in the development of cities, the content of the argument, were demarcated, mutually exclusive stages or were gradual, overlapping stages was critical to responding to the item. Interpretation of certain phrases (i.e., "complex" in "complex divisions of labor") was further critical in whether or not subjects eliminated certain of the options from further consideration.

The directions for the item suggested a potentially advantageous strategy for responding to the item (i.e., drawing a "rough" diagram). Such a strategy, however, was not enumerated in the description/discussion for such items in the Bulletin. Other strategies had been enumerated in the same description/discussion, however, these other strategies were not reiterated in the directions for the item. No rationale for the selective listing of the one strategy in the direc-

(table continues)

Content/Construct Validity

tions for the item, at the exclusion of the other strategies, was provided.

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 1, incorrect responses = 3).

Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement constituted a source of variance in item scores, as two of the incorrect responses were analogous to the "wrong answer for the right reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the length of the argument constituting the basis of the item;
- the extent to which subjects were or were not to assume any information beyond that presented in the argument;
- the style in which the argument was written (e.g., "concrete" versus "abstract");
- prior exposure to and/or familiarity with the type of item.

In conjunction with item GRE/ALR/V/24, these two items were generally parallel in terms of content/construct validity, with the exception of inductive reasoning in the former item (see respective sections for both items). With respect to

(table continues)

internal consistency/discrimination, interindividual differences and intra-individual consistencies were not parallel between the two items (see respective sections for both items). In terms of alternate form/test-retest reliability, the sources of difficulty between the two items were not parallel (see respective sections for both items).

Table 26

GRE/ALR/V/25: Summary of Methodological Inferences

Within-Method Triangulation

The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria:

- The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item.
- The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.

Between-Method Triangulation

The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion:

- (Strategies utilized by subjects in responding to the item were not identified in the literature.)
- (A task analysis relevant to the item was not identified in the literature).
- A review of other studies having utilized various so-called "marker tests" of logical reasoning resulted in the conclusion that few of such studies:

  ... yielded a clear syllogistic or logical reasoning factor. [Certain examinations] tended to load on factors which also included induction tests ... [and certain other tests] tended to load on factors with vocabulary and/or general reasoning tests. This suggests that [logical reasoning] tests do not function similarly [across all subjects and administrations] ... (Ekstrom et al., 1979, p. 36).

tion and discussion of analytical reasoning, presumed to be equivalent to an operational definition.

> Analytical reasoning questions test the ability to understand a given structure of arbitrary relationships among fictitious persons, places, things, or events; to deduce new information from the relationships given; and to assess the conditions used to establish the structure of relationships. .... These relationships are common ones such as temporal order ..., spatial order ..., set membership ..., cause and effect ..., and family relationship ... (p. 19).

No further information relevant to the operational definition or other psychometric attributes of analytical reasoning was provided.

Description of Item GRE/AAR/V/19

The basis of this item was a set of six "conditions", in the form of statements, which described the arrangement of six objects within six locations. Three additional statements, labeled by means of Roman numerals (i.e., I, II, III), were provided and specified the locations of certain of the six objects. The five nonexhaustive options provided for the item consisted of various permutational combinations of the three statements labeled by Roman numerals (e.g., a. I only; b. I and III only). Subjects were to determine which of the options was consistent with the arrangement of the six objects within the six locations as described in the conditions. This item was selected essentially at random from among the items contained in the source examination and was presumed to be parallel to and representative of such other items.

Inferences Relevant to Item GRE/AAR/V/19

Psychometric inferences relevant to item GRE/AAR/V/19 are summarized in Table 27. Methodological inferences relevant to this item are summarized in Table 28.

Logical Reasoning (Kit)

Operational Definition of the Items

Provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition for logical, or deductive, reasoning: "The ability to reason from premise to conclusion, or to evaluate the correctness of a conclusion" (p. 141). No further information relevant to the operational definition or other psychometric attributes of logical reasoning was provided.

Description of the Items

KIT/RL1/1/2.

This item was in the form of a three-sentence syllogism (e.g., No X is Y. All X is Z. Therefore, no X is Z), expressed in "nonsensical" content. Subjects were to assume that the first two statements were "true" and were to determine whether the conclusion expressed in the third statement was consistent with what was termed "good" or "poor" reasoning, given the first two statements. This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to

Table 27

GRE/AAR/V/19: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - general reasoning (see operational definition for item KIT/RG3/1/12); - logical reasoning (see operational definition for item KIT/RL1/1/2); - inductive reasoning (see operational definition for item KIT/I2/1/5); - reading comprehension (i.e., of one of the conditions serving as the basis of the item); - familiarity with the type of item (i.e., "knowing what to look for and how to approach" the item); - efficiency of responding (i.e., "speededness"); - concentrating ability; - capacity of memory (i.e., which arrangements of the objects and locations had been attempted); - consideration of all possible solutions/interpretations. The manner in which one of the conditions on which the item was based was expressed necessitated numerous readings prior to and while responding to the item (i.e., "[object] N is the same ... [distance] ... from [object] M as [object] M is from [object] L"). Provided in the description/discussion relevant to this item in the Bulletin was a caution, advising subjects "... to pay particular attention to function words that describe or limit relationships, such as ONLY, EXACTLY, NEVER, ALWAYS, MUST BE, CANNOT BE, and the like" (p. 19). This precaution was not reiterated in the directions for the item, however, and consideration of the words MUST BE was, in fact, critical for this item. In the item, the words must be were not emphasized (e.g., capital letters, underlining). Within this context, the directions for the item did not advise subjects to consider all possible arrangements of the six objects in the six locations, further critical to responding correctly to the item.

(table continues)

Content/Construct Validity

The multiple response multiple-choice format of this item (i.e., a. I only, b. I and III only) was "annoying" to subjects by virtue of the fact that the response format essentially required subjects to respond to three "true-false" items (i.e., the statements labeled by means of Roman numerals) and then, based on the "true-false" item responses, select a corresponding multiple-choice option.

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 2, incorrect responses = 1, omitted responses = 1).

Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement constituted a source of variance in item scores, as one of the correct responses was analogous to the "right answer for the wrong reason" and both the incorrect and omitted responses were analgous to the "wrong answer for the right reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the clarity with which the conditions on which the item was based were expressed;
- whether or not subjects were advised to consider all possible arrangements of the given objects in given locations (i.e., to attend to words such as MUST BE);
- prior exposure to and/or familiarity with the type of item.

Table 28

GRE/AAR/V/19: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - (Strategies utilized by subjects in responding to the item were not identified in the literature.) - (A task analysis relevant to the item was not identified in the literature.) - A review of other studies having utilized various so-called "marker tests" of analytical, or syllogistic, reasoning resulted in the conclusion that few of such studies: ... yielded a clear syllogistic or logical reasoning factor. [Certain examinations] tended to load on factors which also included induction tests ... [and certain other tests] tended to load on factors with vocabulary and/or general reasoning tests. This suggests that [syllogistic reasoning] tests do not function similarly [across all subjects and administrations] ... (Ekstrom et al., 1979, p. 36).

and representative of such other items.

KIT/RL3/1/9.

This item presented a brief paragraph, consisting of two sentences, as the stem of the item. Five nonexhaustive options were provided and consisted of conclusions which might be drawn from the paragraph. Subjects were to select the conclusion which could be drawn from the paragraph, if no information beyond that provided in the paragraph were assumed. This item was selected essentially at random from the items contained within the source examination and was presumed to be parallel to and representative of such other items.

KIT/RL4/1/4.

This item was on the order of a crytography exercise. Subjects were provided with three three-word phrases which had been "translated" into an artifical language, consisting of letter and symbol characters. The five nonexhaustive options provided for the item consisted of artificial language expressions; subjects were to select the option which corresponded to the phrase constituting the stem of the item. This item was selected from among the items contained within the source examination to represent a "moderate" level of difficulty and was presumed to be parallel to and representative of such other "moderate" level of difficulty items.

KIT/RL4/1/4a.

This item was modified from the former item by the in-

vestigator so as to be posed in a constructed/unrestrictive response format, the former item having been posed in a selected/nonexhaustive response format. Subjects were provided with the identical three three-word phrases which had been translated into the same artificial language as in the former item. An artificial language expression constituted the stem of this item, and subjects were to write the phrase which corresponded to the artificial language expression in the blank space provided. The directions for this item were modified from those of the former item only so as to be consistent with the constructed response format.

Inferences Relevant to the Items

Psychometric inferences relevant to item KIT/RL1/1/2 are summarized in Table 29; methodological inferences relevant to item KIT/RL1/1/2 are summarized in Table 30. Psychometric inferences relevant to item KIT/RL3/1/9 are summarized in Table 31; methodological inferences relevant to item KIT/RL3/1/9 are summarized in Table 32. Psychometric inferences relevant to item KIT/RL4/1/4 are summarized in Table 33; methodological inferences relevant to this item are summarized in Table 34. Psychometric inferences relevant to item KIT/RL4/1/4a are summarized in Table 35; methodological inferences relevant to item KIT/RL4/1/4a are summarized in Table 36.

Table 29

KIT/RL1/1/2: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - familiarity with the type of item (i.e., "knowing how to approach" the item); - terminology of formal logic (i.e., "good" or "poor" reasoning); - general reasoning (see operational definition for item KIT/RG3/1/12); - inductive reasoning (see operational definition for item KIT/I2/1/5); - consideration of all possible solutions/interpretations. The directions for the item included the term "syllogism", without defining or otherwise explaining what was meant by the term. Subjects unfamiliar with the term were unable to discern what was meant by the term merely from reading the directions and found the term distracting. The criteria to be utilized in assessing the conclusion represented by the third statement was whether "good" or "poor" reasoning were exhibited. However, the terms "good" and "poor" reasoning were never defined or otherwise explained. The practice items included on the cover sheet of the examination, in actual administration, indicated the correct responses to the practice items, however, no explanations were provided relevant to the practice items. In the absence of such explanations, the presumed purpose of providing practice items (i.e., ensuring that subjects comprehended the task posed by the item) was only partially accomplished. The directions provided for scoring in the source examination specified that examination scores would be "corrected for guessing" (i.e., scores equal to the number of items marked correctly minus the number of items marked incorrectly). However, the rationale for such a scoring procedure was not provided in the description/discussion for the examination in the Manual.

(table continues)

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 0, incorrect responses = 4).

Aptitudes did not constitute a source of variance in item scores, although interindividual differences in the aptitudes previously delineated were exhibited.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- whether or not the directions specified that subjects were to consider all possible solutions/interpretations (i.e., the clarity with which the task posed by the item was specified);
- prior exposure to and/or familiarity with the type of item.

(See also corresponding section for item KIT/RL4/1/4a.)

Table 30

KIT/RL1/1/2: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - (Strategies utilized by subjects in responding to the item were not identified in the literature.) - (A task analysis relevant to the item was not identified in the literature.) - A review of other studies having utilized various so-called "marker tests" of logical reasoning resulted in the conclusion that few of such studies: ... yielded a clear syllogistic or logical reasoning factor. [Certain examinations] tended to load on factors which also included induction tests ... [and certain other tests] tended to load on factors with vocabulary and/or general reasoning tests. This suggests that [logical reasoning] tests do not function similarly [across all subjects and administrations] ... (Ekstrom et al., 1979, p. 36).

Table 31

KIT/RL3/1/9: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - reading comprehension; - vocabulary (i.e., relevant to geology, the content of the paragraph on which the item was based); - consideration of all possible solutions/interpretations; - susceptibility or resistance to interference from knowledge previously acquired; - general reasoning (see operational definition for item KIT/RG3/1/12). The directions for the item did not explicitly or implicitly state that more than one solution/interpretation was possible of the paragraph on which the item was based, which was critical to correctly responding to the item. Without assuming any information beyond that provided in the paragraph on which the item was based, whether or not the correct, keyed response is, in fact, the correct response and the only correct response is indeterminate. The directions provided for scoring in the source examination specified that examination scores would be "corrected for guessing" (i.e., scores equal to the number of items marked correctly minus a fraction of the number of items marked incorrectly). However, the rationale for such a scoring procedure was not provided in the description/discussion for the examination in the Manual. Furthermore, nowhere in the directions for scoring the examination was the fraction to be utilized in the "correction for guessing" formula specified.

(table continues)

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 1, incorrect responses = 3), assuming that the correct, keyed response is the only correct response.

Aptitudes consituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement constituted a source of variance in item scores, as two incorrect responses were analogous to the "wrong answer for the right reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the level of reading comprehension and vocabulary required for responding to the item;
- the susceptibility or resistance to interference from knowledge previously acquired.

Item scores as well as examination scores would vary within and between both investigators and studies dependent upon the fraction utilized in the "correction for guessing" formula.

(See also corresponding section for item KIT/RL4/1/4a.)

Table 32

KIT/RL3/1/9: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - (Strategies utilized by subjects in responding to the item were not identified in the literature.) - (A task analysis relevant to the item was not identified in the literature.) - A review of other studies having utilized various so-called "marker tests" of logical reasoning resulted in the conclusion that few of such studies: ... yielded a clear syllogistic or logical reasoning factor. [Certain examinations] tended to load on factors which also included induction tests ... [and certain other tests] tended to load on factors with vocabulary and/or general reasoning tests. This suggests that [logical reasoning] tests do not function similarly [across all subjects and administrations] ... (Ekstrom et al., 1979, p. 36).

Table 33

KIT/RL4/1/4: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- general reasoning (see operational definition for item KIT/RG3/1/12);
- inductive reasoning (see operational definition for item KIT/I2/1/5);
- perceptual speed (see operational definition for item KIT/P2/1/10);
- integrative processes (see operational definition for item KIT/IP1/1/9);
- associative memory (see operational definition for item KIT/MA3/1/-) and/or memory span (i.e., "The ability to recall a number of distinct elements for immediate reproduction"; Ekstrom et al., 1976b, p. 101);
- reading comprehension (i.e., of the directions provided for the item);
- consideration of all possible solutions/interpretations;
- familiarity with the type of item.

The directions for the item contained two sentences of explanation concerning the order of the words and symbols in the phrases and artificial language expressions. Had an illustration/example of what was meant by the two sentences been provided, perhaps subjects would have more readily understood the sentences, without repeated readings prior to and while responding to the item.

One of the artificial language expressions serving as the basis for the item contained a typographical error which had not been completely "erased" (i.e., the typographical error "showed through" the correction). The typographical error served as a source of confusion and/or distraction for subjects who presumed, initially, that the extraneous mark was intended as part of the artificial language expression in which it appeared.

The practice items included on the cover sheet of the exami-

(table continues)

Content/Construct Validity

nation, in actual administration, indicated the correct responses to the practice items, however, no explanations were provided relevant to the practice items. In the absence of such explanations, the presumed purpose of providing practice items (i.e., ensuring that subjects comprehended the task posed by the item) was only partially accomplished.

The directions provided for scoring in the source examination specified that examination scores would be "corrected for guessing" (i.e., scores equal to the number of items marked correctly minus a fraction of the number of items marked incorrectly). However, the rationale for such a scoring procedure was not provided in the description/discussion for the examination in the Manual. Furthermore, nowhere in the directions for scoring the examination was the fraction to be utilized in the "correction for guessing" formula specified.

Internal Consistency/Discrimination

The item responses of the subjects revealed no interindividual differences in terms of item scores (i.e., correct responses = 4, incorrect responses = 0).

Aptitudes did not constitute a source of variance, although interindividual differences in the aptitudes previously delineated were exhibited.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

(table continues)

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the number of words and symbols common to the three phrases and artificial language expressions serving as the basis of the item;
- the response format in which the item was posed (e.g., selected/nonexhaustive versus selective/exhaustive).

Item scores as well as examination scores would vary within and between both investigators and studies dependent upon the fraction utilized in the "correction for guessing" formula.

Not all parallel items in the source examination may constitute independent measures. In actual administration of the source examination, from three to six items are based on a single set of phrase/artificial language expressions. For certain sets of the three to six items, items within that set may be responded to by application or transfer of translations performed in preceding items of that same set.

(See also corresponding section for item KIT/RL4/1/4a).

Table 34

KIT/RL4/1/4: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.

Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - (Strategies utilized by subjects in responding to the item were not identified in the literature.) - (A task analysis relevant to the item was not identified in the literature.) - A review of other studies having utilized various so-called "marker tests" of logical reasoning resulted in the conclusion that few of such studies: ... yielded a clear syllogistic or logical reasoning factor. [Certain examinations] tended to load on factors which also included induction tests ... [and certain other tests] tended to load on factors with vocabulary and/or general reasoning tests. This suggests that [logical reasoning] tests do not function similarly [across all subjects and administrations] ... (Ekstrom et al., 1979, p. 36).

Table 35

KIT/RL4/1/4a: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- general reasoning (see operational definition for item KIT/RG3/1/12);
- inductive reasoning (see operational definition for item KIT/I2/1/5);
- perceptual speed (see operational definition for item KIT/P2/1/10);
- integrative processes (see operational definition for item KIT/P2/1/10);
- associative memory (see operational definition for item KIT/MA3/1/-) and/or memory span (i.e., "The ability to recall a number of distinct elements for immediate reproduction"; Ekstrom et al., 1976b, p. 101).
- reading comprehension (i.e., of the directions provided for the item);
- consideration of all possible solutions/interpretations;
- familiarity with the type of item.

The directions for the item contained two sentences of explanation concerning the order of the words and symbols in the phrases and artificial language expressions. Had an illustration/example of what was meant by the two sentences been provided, perhaps subjects would have more readily understood the sentences, without repeated reading prior to and while responding to the item.

One of the artificial language expressions serving as the basis for the item contained a typographical error which had not been completely "erased" (i.e., the typographical error "showed through" the correction). The typographical error served as a source of confusion and/or distraction for subjects who presumed, initially, that the extraneous mark was intended as part of the artificial language expression in which it appeared.

The practice items included on the cover sheet of the exami-

(table continues)

Content/Construct Validity

nation, in actual administration, indicated the correct responses to the practice items, however, no explanations were provided relevant to the practice items. In the absence of such explanations, the presumed purpose of providing practice items (i.e., ensuring that subjects comprehended the task posed by the item) was only partially accomplished).

Internal Consistency/Discrimination

The item responses of the subjects revealed no interindividual differences in terms of item scores (i.e., correct responses = 4, incorrect responses = 0).

Aptitudes did not constitute a source of variance, although interindividual differences in the aptitudes previously delineated were exhibited.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the number of words and symbols common to the three phrases and artificial language expressions serving as the basis of the item;
- the response format in which the item was posed (e.g., selected/nonexhaustive versus selected/exhaustive versus constructed).

Not all parallel items in the source examination may constitute independent measures. In actual administration of the

(table continues)

source examination, from three to six items are based on a single set of phrase/artificial language expressions. For certain sets of the three to six items, items within that set may be responded to by application or transfer of translations performed in preceding items of that same set.

In conjunction with items KIT/RL1/1/2, KIT/RL3/1/9, and KIT/RL4/1/4, these four items were not parallel in terms of content/construct validity (see respective sections for all four items), except for this item and item KIT/RL4/1/4. With respect to internal consistency/discrimination, interindividual differences and intra-individual consistencies were not parallel across the four items (see respective sections for all four items). In terms of alternate form/test-retest reliability, the sources of difficulty across the four items were not parallel (see respective sections for all four items).

Table 36

KIT/RL4/1/4a: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - (Strategies utilized by subjects in responding to the item were not identified in the literature.) - (A task analysis relevant to the item was not identified in the literature.) - A review of other studies having utilized various so-called "marker tests" of logical reasoning resulted in the conclusion that few of such studies: ... yielded a clear syllogistic or logical reasoning factor. [Certain examinations] tended to load on factors which also included induction tests ... [and certain other tests] tended to load on factors with vocabulary and/or general reasoning tests. This suggests that [logical reasoning] tests do not function similarly [across all subjects and administrations] ... (Ekstrom et al., 1979, p. 36).

Inductive Reasoning

Operational Definition of the Items

Provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition for inductive reasoning: "This factor identifies the kinds of reasoning abilities involved in forming and trying out hypotheses that will fit a set of data" (p. 79). No further information relevant to the operational definition or other psychometric attributes of inductive reasoning was provided.

Description of the Items

KIT/I2/1/5.

The stem of this item consisted of five rows of "dashes" and "spaces" (e.g., --- ------ - --). Within each of the first four rows, an "x" had been substituted into the row (e.g., --- ---x- --). In the fifth row, five Arabic numerals had been substituted into the row and represented the five nonexhaustive options for the item. Subjects were to determine what "rule" had governed the placement of the "x's" in the first four rows and, by extending that rule, were to determine which of the five options corresponded to where the "x" would be placed in the fifth row. This item was selected from among the items contained within the source examination to represent a "marked" level of difficulty and was presumed to be parallel to and representative of such other "marked"-level-of-difficulty items.

KIT/I2/1/5a.

This item was modified from the former item by the investigator so as to be posed in a selected/exhaustive response format, the former item having been posed in a selected/nonexhaustive response format. Subjects were provided with the same five rows of "dashes" and "spaces" as in the former item, with the exception that in the fifth row, "dashes" had been substituted back into the row to replace the Arabic numerals representing the five options in the former item (i.e., the fifth row consisted of "dashes" and "spaces" only). Subjects were to determine whether or not a "rule" governed the placement of the "x's" in the first four rows. If so, by extending that rule, subjects were to indicate where the "x" would be placed in the fifth row by drawing an "x" through the corresponding "dash" or "space". If not, subjects were to indicate that no rule appeared to govern the placement of the "x's". The directions for this item were modified from those of the former item only so as to be consistent with the selected/exhaustive response format.

KIT/I3/1/7.

This item consisted of two groups of three figures; the figures were composed of line and circle patterns or designs. Subjects were to determine what features were common to the three figures constituting the first group, what features were common to the three figures constituting the second group, and what features differentiated the three figures of

the first group, collectively, from the three figures of the second group, collectively. On the basis of such features, subjects were to assign to either the first or the second group each of eight figures presented as "unknowns". This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

Inferences Relevant to the Items

Psychometric inferences relevant to item KIT/I2/1/5 are summarized in Table 37; methodological inferences relevant to this item are summarized in Table 38. Psychometric inferences relevant to item KIT/I2/1/5a are summarized in Table 39; methodological inferences relevant to this item are summarized in Table 40. Psychometric inferences relevant to item KIT/I3/1/7 are summarized in Table 41; methodological inferences relevant to this item are summarized in Table 42.

Associative Memory

Operational Definition of Item KIT/MA3/1/-

Provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition for associative memory: "The ability to recall one part of a previously learned but otherwise unrelated pair of items when the other part of the pair is presented" (p. 93). No further information relevant to the

Table 37

KIT/I2/1/5: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- general reasoning (see operational definition for item KIT/RG3/1/12);
- logical reasoning (see operational definition for item KIT/RL1/1/2);
- flexibility of closure (see operational definition for item KIT/CF1/1/12);
- pattern recognition;
- efficiency of responding (i.e., "speededness");
- familiarity with the type of item.

The directions provided for the item specified that "... any kind of relation or rule to explain the position of the x's" was possible. However, seemingly the "rules" governing the placement of the "x's" within the rows for all items in the source examination were of a "quantitative" type (e.g., first dash in the next to the last group of dashes in all five rows). Given the lack of specificity or ambiguity in the directions for the item, one subject utilized a "symbolic rule" as the basis for responding to the item. However, given the range of potential "rules" implied in the directions, perhaps the subject's item response was "justified" as a correct response.

The directions provided for scoring in the source examination specified that examination scores would be "corrected for guessing" (i.e., scores equal to the number of items marked correctly minus a fraction of the number of items marked incorrectly). However, the rationale for such a scoring procedure was not provided in the description/discussion of the examination in the Manual. Furthermore, nowhere in the directions provided for scoring the examination was the fraction to be utilized in the "correction for guessing" specified.

(table continues)

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 0, incorrect responses = 3, omitted responses = 1).

Aptitudes did not constitute a source of variance in item scores, although interindividual differences in the aptitudes previously delineated were exhibited.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the complexity of the "rule" governing the placement of the "x's" in the item;
- the response format of the item (e.g., selected/nonexhaustive versus selected/exhaustive).

Item scores as well as examination scores would vary within and between investigators and studies dependent upon the fraction utilized in the "correction for guessing" formula.

(See also corresponding section for item KIT/I3/1/7.)

Table 38

KIT/I2/1/5: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - Strategies utilized by subjects in responding to the item paralleled those described in the literature (Carroll, 1976; Pellegrino and Glaser, 1979). - The responses of the subjects to the item paralleled the task analysis identified in the literature (Pellegrino and Glaser, 1979). - The responses of the subjects were consistent with evidence from the literature that inductive reasoning may not constitute a univariate factor (Green et al., 1953; French, 1957; 1965; Pellegrino and Glaser, 1979; Nunnally, 1978; Sternberg, 1977; Ekstrom et al., 1976b).

Table 39

KIT/I2/1/5a: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- general reasoning (see operational definition for item KIT/RG3/1/12);
- logical reasoning (see operational definition for item KIT/RL1/1/2);
- flexibility of closure (see operational definition for item KIT/CF1/1/12);
- pattern recognition;
- efficiency of responding (i.e., "speededness");
- familiarity with the type of item.

The directions provided for the item specified that "... any kind of relation or rule to explain the position of the x's" was possible. However, seemingly the "rules" governing the placement of the "x's" within the rows for all items in the source examination were of a "quantitative" type (e.g., first dash in the next to the last group of dashes in all five rows). Given the lack of specificity or ambiguity in the directions for the item, one subject utilized a "symbolic" rule as the basis for responding to the item. However, given the range of potential "rules" implied in the directions, perhaps the subject's item response was "justified" as a correct response.

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 2, incorrect responses = 1, omitted responses = 1).

Aptitudes did not constitute a source of variance in item scores, although interindividual differences in the aptitudes previously delineated were exhibited.

(table continues)

Internal Consistency/Discrimination

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement constituted a source of variance in item scores, as the two correct responses were analogous to the "right answer for the wrong reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the complexity of the "rule" governing the placement of the "x's" in the item;
- the response format of the item (e.g., selected/nonexhaustive versus selected/exhaustive).

(See also corresponding section for item KIT/I3/1/7.)

Table 40

KIT/I2/1/5a: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - Strategies utilized by subjects in responding to the item paralleled those described in the literature (Carroll, 1976; Pellegrino and Glaser, 1979). - The responses of the subjects to the item paralleled the task analysis identified in the literature (Pellegrino and Glaser, 1979). - The responses of the subjects were consistent with evidence from the literature that inductive reasoning may not constitute a univariate factor (Green et al., 1953; French, 1957, 1965; Pellegrino and Glaser, 1979; Nunnally, 1978; Sternberg, 1977; Ekstrom et al., 1976b).

Table 41

KIT/I3/1/7: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - general reasoning (see operational definition for item KIT/RG3/1/12); - logical reasoning (see operational definition for item KIT/RL1/1/2); - flexibility of closure (see operational definition for item KIT/CF1/1/12); - pattern recognition/concept formation; - perceptual speed (see operational definition for item KIT/P2/1/10); - speed of closure (i.e., "The ability to unite an apparently disparate perceptual field into a single concept"; Ekstrom et al., 1979, p. 25); - familiarity with the type of item. The directions provided for the item were not sufficiently explicit that the three figures in each of the two groups were to be considered collectively in order to determine the features of the groups of figures which were common and different. Given the lack of specificity or ambiguity in the directions and the unfamiliarity of the subjects with the type of item, all four subjects attempted to match each of the "unknown" figures with individual figures in the first and second groups (i.e., a one-to-one correspondence). The directions provided for scoring in the source examination specified that examination scores would be "corrected for guessing" (i.e., scores equal to the number of items marked correctly minus a fraction of the number of items marked incorrectly). However, the rationale for such a scoring procedure was not provided in the description/discussion of the examination in the Manual. Furthermore, nowhere in the directions for scoring the examination was the fraction to be utilized in the "correction for guessing" formula specified.

(table continues)

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., number of figures correctly assigned to groups = 8, 5, 7, 6).

Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement constituted a source of variance in item scores, as certain of the correct item responses for all four subjects were analogous to the "right answer for the wrong reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the complexity of the figures serving as the basis of the item;
- the number of relevant versus irrelevant features contained in the figures serving as the basis for the item;
- the response format of the item (e.g., selected/nonexhaustive versus selected/exhaustive).

In conjunction with items KIT/I2/1/5 and KIT/I2/1/5a, these three items were not parallel in terms of content/construct validity (see respective sections for all three items), except for the former two items. With respect to internal consistency/discrimination, interindividual differences and intra-individual consistencies were not parallel across the three items (see respective sections for all three items). In terms of alternate form/test-retest reliability, the sources of difficulty across the three items were relatively parallel (see respective sections for all three items).

Table 42

KIT/I3/1/7: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - Strategies utilized by subjects in responding to the item paralleled those described in the literature (Carroll, 1976; Pellegrino and Glaser, 1979). - The responses of the subjects to the item paralleled the task analysis identified in the literature (Pellegrino and Glaser, 1979). - The responses of the subjects were consistent with evidence from the literature that inductive reasoning may not constitute a univariate factor (Green et al., 1953; French, 1957, 1965; Pellegrino and Glaser, 1979; Nunnally, 1978; Sternberg, 1977; Ekstrom et al., 1976b).

operational definition or other psychometric attributes of associative memory was provided.

Description of Item KIT/MA3/1/-

The stem of the item consisted of a list of fifteen pairs of first and last names. After studying the list, subjects were to be presented with a second list which consisted of only the last names in a different order from that of the first list, and were to write in the blank line preceding each last name the first name which had been paired with that last name. This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

Inferences Relevant to Item KIT/MA3/1/-

Psychometric inferences relevant to this item are summarized in Table 43. Methodological inferences relevant to this item are summarized in Table 44.

Spatial Visualization

Operational Definition of Item KIT/VZ3/2/8

Provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition for spatial visualization: "The ability to manipulate or transform the image of spatial patterns into other arrangements" (p. 173). No further information relevant to the operational definition or other psychometric

Table 43

KIT/MA3/1/-: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - concentrating ability; - susceptibility or resistance to interference from knowledge previously acquired (i.e., names of other individuals); - efficiency of responding (i.e., "speededness"). The directions for the item implicitly inferred that after studying the list of names on the first page, subjects would not be permitted to refer back to the first page. Such information was not, however, explicitly stated. The directions for the item stated that "[e]ven if you are not sure of the correct answer to a question, it will be to your advantage to guess". No rationale or other explanation was provided in the description/discussion for the examination in the Manual.
Internal Consistency/Discrimination
The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., number of first names correctly listed = 3, 4, 11, 0). Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated. Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited. Random errors of measurement did not constitute a source of

(table continues)

Internal Consistency/Discrimination

variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the response format of the item (e.g., selected versus constructed).

Table 44

KIT/MA3/1/-: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - Strategies utilized by subjects in responding to the item paralleled those described in the literature (Bower and Hilgard, 1981; Frederiksen, 1969; Carroll, 1976). - The responses of the subjects to the item paralleled the task analyses identified in the literature (Bower and Hilgard, 1981; Frederiksen, 1969). - The responses of subjects to the item were consistent with the seeming consensus that "[l]arge individual ... differences can be obtained in [memory] task[s]" (Pellegrino and Glaser, 1979, p. 70).

attributes of spatial visualization was provided.

Description of Item KIT/VZ3/2/8

This item was of the type described as "mental paper folding". The basis for the item consisted of two drawings of a three-dimensional geometric figure. The first drawing was that of the figure in an "unfolded" state, representing a "pattern" of the figure; various edges of the "unfolded" drawing had been labeled with Arabic numerals. The second drawing was that of the figure in a "folded" state representing a solid, opaque object; the visible edges of the "folded" drawing had been labeled with letters. For the five numbered edges of the "unfolded" drawing indicated, subjects were to write the letter labeling the edge of the "folded" drawing which corresponded to that numbered edge in the blank spaces provided. This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

Inferences Relevant to Item KIT/VZ3/2/8

Psychometric inferences relevant to this item are summarized in Table 45. Methodological inferences relevant to this item are summarized in Table 46.

Perceptual Speed

Operational Definition of Item KIT/P2/1/10

Provided in the Manual for Kit of Factor-Referenced

Table 45

KIT/VZ3/2/8: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - visual memory (i.e., "The ability to remember the configuration, location, and orientation of figural material"; Ekstrom et al., 1976b, p. 109); - perceptual speed (see operational definition for item KIT/P2/1/10); - flexibility of closure (see operational definition for item KIT/CF1/1/12); - spatial orientation (i.e., "The ability to perceive spatial patterns or to maintain orientation with respect to objects in space"; Ekstrom et al., 1976b, p. 149); - efficiency of responding (i.e., "speededness"); - capacity of memory (i.e., to retain results of serial, consecutive folding operations). The manner in which the edges of the "folded" drawing were labeled was confusing to subjects, in that subjects were uncertain whether the labels referred to the edges of the drawing or to the planes of the drawing. That the labels referred to the edges of the drawing was stated unambiguously in the directions for the item, however, was perhaps not sufficiently emphasized (e.g., capital letters, underlining). The directions provided for scoring in the source examination specified that examination scores would be "corrected for guessing" (i.e., scores equal to the number of items marked correctly minus a fraction of the number of items marked incorrectly). However, the rationale for such a scoring procedure was not provided in the description/discussion of the examination in the Manual. Furthermore, nowhere in the directions for scoring the examination was the fraction to be utilized in the "correction for guessing" formula specified.

(table continues)

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of items scores (i.e., number of edges correctly identified = 4, 2, 5, 1).

Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement constituted a source of variance in item scores, as certain of the item responses for three subjects were analogous to the "right answer for the wrong reason".

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the complexity of the drawing depicted and which edges of the drawing were to be identified;
- the response format of the item (e.g., selected/nonexhaustive versus selected/exhaustive);
- prior exposure to and/or familiarity with the type of item.

Item scores as well as examination scores would vary within and between both investigators and studies dependent upon the fraction utilized in the "correction for guessing" formula.

Table 46

KIT/VZ3/2/8: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - Strategies utilized by subjects in responding to the item paralleled those described in the literature (Pellegrino and Glaser, 1979; Ekstrom et al., 1976b; Nunnally, 1978). - (A task analysis relevant to this item was not identified in the literature). - The results of various other studies have suggested that spatial visualization may represent a more difficult form of perceptual speed and spatial orientation and may consist of visual memory and flexibility of closure components as well (see operational definitions in the Content/Construct Validity section of this item) (Pellegrino and Glaser, 1979; Ekstrom et al., 1976b, Nunnally, 1978).

Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition for perceptual speed: "Speed in comparing figures or symbols, scanning to find figures or symbols, or carrying out other very simple tasks involving visual perception" (p. 123). No further information relevant to the operational definition or other psychometric attributes of perceptual speed was provided.

Description of Item KIT/P2/1/10

The stem of the item consisted of two series of 12 Arabic numerals, one series to the right and one series to the left of a blank line. Subjects were to compare the two series of numerals and place an "x" on the blank line if the two series of numerals were not identical and not place an "x" on the blank line if the two series of numerals were identical. This item was selected from among the items contained within the source examination to represent a "long" series of numerals and was presumed to be parallel to and representative of such other "long"-series-of-numeral items.

Inferences Relevant to Item KIT/P2/1/10

Psychometric inferences relevant to this item are summarized in Table 47. Methodological inferences relevant to this item are summarized in Table 48.

Flexibility of Closure

Operational Definition for Item KIT/CF1/1/12

Provided in the Manual for Kit of Factor-Referenced

Table 47

KIT/P2/1/10: Summary of Psychometric Inferences

Content/Construct Validity

The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item:

- familiarity with the numerals constituting the stem of the item (i.e., enabling "immediate recognition" versus "analysis");
- auditory discrimination/perception (i.e., "hearing" differences between the series of numerals when reading aloud or to one's self);
- memory span (i.e., "The ability to recall a number of distinct elements for immediate reproduction"; Ekstrom et al., 1976b, p. 101).

Given the manner in which subjects were to indicate their item responses (e.g., not to place an "x" on the blank line if the two series of numerals were identical), scores for subjects, in actual administration of the examination, who were unable to complete the examination in the allotted time would be inflated dependent on whether all items were scored according to the above criterion (i.e., dependent upon the number of items not attempted for which the correct, keyed response was "no x" on the blank line).

The directions provided for scoring in the source examination specified that examination scores would be "corrected for guessing" (i.e., scores equal to the number of items marked correctly minus the number of items marked incorrectly). However, the rationale for such a scoring procedure was not provided in the description/discussion of the examination in the Manual.

Internal Consistency/Discrimination

The item responses of the subjects revealed no interindividual differences in terms of item scores (i.e., correct re-

(table continues)

Internal Consistency/Discrimination

sponses = 4, incorrect responses = 0).

Aptitudes did not constitute a source of variance in item scores, although interindividual differences in the aptitudes previously delineated were exhibited.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the length of the series of numerals constituting the stem of the item (e.g., 32681 versus 4832709285?).
- the complexity of the stimulus constituting the stem of the item (e.g., series of numerals versus symbols [#/@!&+?] versus figures [pictures of faces, houses, other line drawings]).

Item scores as well as examination scores would vary within and between both investigators and studies dependent upon the manner in which unattempted items for which the correct, keyed responses was "no x" were scored.

Table 48

KIT/P2/1/10: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - Strategies utilized by subjects in responding to the item paralleled those described in the literature (Bower and Hilgard, 1981; Frederiksen, 1969). - (A task analysis relevant to this item was not identified in the literature.)

Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition for flexibility of closure: "The ability to hold a given visual percept or configuration in mind so as to disembed it from other well defined perceptual material" (p. 19). No further information relevant to the operational definition or other psychometric attributes of flexibility of closure was provided.

Description of Item KIT/CF1/1/12

This item was of the type variously referred to as "hidden figures" or "embedded figures". The item was based on a geometric, line drawing contained within the boundaries of a square. Five nonexhaustive options were provided for the item, each option consisting of a geometric, line drawing. Subjects were to determine which of the drawings provided as the options for the item was contained within the drawing serving as the basis for the item. This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

Inferences Relevant to Item KIT/CF1/1/12

Psychometric inferences relevant to this item are summarized in Table 49. Methodological inferences relevant to this item are summarized in Table 50.

Table 49

KIT/CF1/1/12: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition of the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - perceptual speed (see operational definition for item KIT/P2/1/10); - concentrating ability; - visual memory (i.e., "The ability to remember the configuration, location, and orientation of figural material"; Ekstrom et al., 1976b, p. 109); - efficiency of responding (i.e., "speededness"). The directions provided for scoring in the source examination specified that examination scores would be "corrected for guessing" (i.e., scores equal to the number of items marked correctly minus a fraction of the number of items marked incorrectly). However, the rationale for such a scoring procedure was not provided in the description/discussion of the examination in the Manual. Furthermore, nowhere in the directions for scoring the examination was the fraction to be utilized in the "correction for guessing" formula specified.
Internal Consistency/Discrimination
The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., correct responses = 1, incorrect responses = 1, omitted responses = 2). Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated. Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

(table continues)

Internal Consistency/Discrimination

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the response format of the item (e.g., selected/nonexhaustive versus selected/exhaustive);
- whether, in addition to indicating which of the drawings provided as options was contained within the one serving as the basis of the item, subjects were required to trace the outline of the drawing option contained within the drawing serving as the basis for the item.

Item scores as well as examination scores would vary within and between both investigators and studies dependent upon the fraction utilized in the "correction for guessing" formula.

Table 50

KIT/CF1/1/12: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - Strategies utilized by subjects in responding to the item paralleled those described in the literature (French, 1965; Pellegrino and Glaser, 1979). - (A task analysis relevant to this item was not identified in the literature.) - The conclusions of various other studies have suggested that flexibility of closure represents a not well-defined factor consisting of multiple components or sources of variance yet to be adequately delineated (French, 1965; Pellegrino and Glaser, 1979; Ekstrom et al., 1976b, 1979).

Integrative Processes

Operational Definition for Item KIT/IP1/1/9

Provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition for integrative processes: "The ability to keep in mind simultaneously or to combine several conditions, premises, or rules in order to produce a correct response" (p. 87). No further information relevant to the operational definition or other psychometric attributes of integrative processes was provided.

Description of Item KIT/IP1/1/9

The stem of this item consisted of a question describing a date on a calendar (e.g., What is the fourth Tuesday ... ?). From the five exhaustive options provided for the item, subjects were to select the date which corresponded to that described in the item. For determining the date described in the item, subjects were also provided with a calendar reproduced on a separate sheet of paper. In determining the date described in the item, subjects were supposed to consider seven "conditions" included as part of the directions for the item (e.g., "[a] circled [date] is a holiday"). This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to and representative of such other items.

Inferences Relevant to Item KIT/IP1/1/9

Psychometric inferences relevant to the item are summarized in Table 51. Methodological inferences relevant to this item are summarized in Table 52.

Flexibility of Use

Operational Definition for Item KIT/XU3/1/2

Provided in the Manual for Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976b) was the following operational definition for flexibility of use: "The mental set necessary to think of different uses for objects" (p. 197). No further information relevant to the operational definition or other psychometric attributes of flexibility of use was provided.

Description of Item KIT/XU3/1/2

A list of seven "things" was provided as the basis for the item. Subjects were to form a maximum of 10 groups, utilizing as a criterion the attributes common between and among the seven "things". Each group was to contain a minimum of three "things". For each resultant group of "things" formed, subjects were to list the letters labeling the "things" on the blank lines provided in a "group" column and to list the reason for having formed the group on the blank lines provided in a "reason" column. This item was selected essentially at random from among the items contained within the source examination and was presumed to be parallel to

Table 51

KIT/IP1/1/9: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - reading comprehension (i.e., of the stem of the item); - perceptual speed (see operational definition for item KIT/P2/1/10); - distractibility; - attention to detail (i.e., counting days in the calendar); - general reasoning (see operational definition for item KIT/RG3/1/12). The directions provided for scoring in the source examination specified that examination scores would be "corrected for guessing" (i.e., scores equal to the number of items marked correctly minus a fraction of the number of items marked incorrectly). However, the rationale for such a scoring procedure was not provided in the description/discussion of the examination in the Manual. Furthermore, nowhere in the directions for scoring the examination was the fraction to be utilized in the "correction for guessing" formula specified.
Internal Consistency/Discrimination
The item responses of the subjects revealed no interindividual differences in terms of item responses (i.e., correct responses = 0, incorrect responses = 4). Aptitudes did not constitute a source of variance in item scores, although interindividual differences in the aptitudes previously delineated were exhibited. Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies

(table continues)

Internal Consistency/Discrimination

were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the number of conditions to be considered relative to the date described in the stem of the item;
- the extent to which reading comprehension of the stem of the item was required to "interpret" or "translate" the date described in the stem of the item.

Item scores as well as examination scores would vary within and between both investigators and studies dependent upon the fraction utilized in the "correction for guessing" formula.

Table 52

KIT/IP1/1/9: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - (Strategies utilized by subjects in responding to the item were not identified in the literature.) - (A task analysis relevant to this item was not identified in the literature.) - A review of the results of the calibration procedures utilized in establishing the source examination as a so-called "marker test" of integrative processes resulted in the conclusion that: "The integrative processes factor seemed to be somewhat indistinct and difficult to separate from some of the reasoning factors" (Ekstrom et al., 1979).

and representative of such other items.

Inferences Relevant to Item KIT/XU3/1/2

Psychometric inferences relevant to item KIT/XU3/1/2 are summarized in Table 53. Methodological inferences relevant to the item are summarized in Table 54.

Table 53

KIT/XU3/1/2: Summary of Psychometric Inferences

Content/Construct Validity
The operational definition for the item did not sufficiently acknowledge the following sources of variance as determinants in responding to the item: - experience/acculturation and/or vocabulary; - capacity of memory (i.e., retention of specifications in the directions, which "things" had been listed in which "groups"); - ideational fluency (see operational definition for item KIT/FI3/2/-); - expressional fluency (see operational definition for item KIT/FE1/2/18); - general reasoning (see operational definition for item KIT/RG3/1/12); - inductive reasoning (see operational definition for item KIT/I2/1/5); - hierarchical clustering/chunking of semantic memory; - efficiency of responding (i.e., "speededness"). The directions provided for the item were lengthy and included numerous specifications to be considered in forming "groups of things", thus numerous readings of the directions were required prior to and while responding to the item. In actual administration of the source examination, the directions for the examination would have been provided on the equivalent of an examination booklet cover sheet. Given the length of the directions and the specifications of the directions, subjects would be required to "flip back and forth" between the cover sheet and the items in order to refer to the directions. The extent to which such "flipping back and forth" would be distracting and/or time-consuming is indeterminate. The directions provided for the item specified that the same group of "things" could not be listed more than once, even if the reason for the grouping were changed. However, the directions did not specify whether or not the same reason

(table continues)

Content/Construct Validity

could be listed for more than one group of "things". Inquiries from subjects to this effect could not be addressed from merely reading the directions.

The directions provided for scoring in the source examination contradicted one specification which had been included in the directions for the item. The directions for scoring further included one additional criterion for assessing the "correctness" of subjects' responses which had not been included in the directions for the item. The directions for scoring provided no further criteria or guidelines for assessing the "correctness" of subjects' responses (e.g., whether to give credit for a reason listed which was not "accurate").

Internal Consistency/Discrimination

The item responses of the subjects revealed interindividual differences in terms of item scores (i.e., number of "correct" groups listed = 6, 10, 6, 4).

Aptitudes constituted a source of variance in item scores, as interindividual differences in item scores were attributable to the aptitudes previously delineated.

Strategies did not constitute a source of variance in item scores, although interindividual differences in strategies were exhibited.

Random errors of measurement did not constitute a source of variance in item scores, as no random errors of measurement were exhibited.

Alternate Form/Test-Retest Reliability

Parallel items would vary in difficulty dependent upon:

- the familiarity or unfamiliarity of the "things" listed

(table continues)

from which subjects were to form groups;
- the extent to which the groups formed from the list of "things" were required to emphasize the "quality" of ideas as opposed to the "quantity" of ideas.

Item scores as well as examination scores would vary within and between both investigators and studies dependent upon the criteria utilized in assessing the "correctness" of item responses of the subjects.

Table 54

KIT/XU3/1/2: Summary of Methodological Inferences

Within-Method Triangulation
The internal validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criteria: - The manner in which subjects anticipated responding to the item paralleled the manner in which subjects actually responded to the item. - The responses of the subjects to the various aspects of the nonschedule standardized interview revealed inter-individual differences in content and comprehensiveness.
Between-Method Triangulation
The external validity of the data source, data collection, and data analysis components of the present study were considered supported by the following criterion: - (Strategies utilized by subjects in responding to the item were not identified in the literature.) - (A task analysis relevant to this item was not identified in the literature.) - A review of the results of the calibration procedures utilized in establishing the source examination as a so-called "marker test" of flexibility of use resulted in the conclusion that flexibility of use could not be distinguished categorically from other measures of semantic and figural fluency, flexibility, and/or originality (Ekstrom et al., 1976b, 1979).

Discussion

Prerequisite and prior to assessment of the psychometric inferences derived in the present study is an assessment of the methodological inferences derived in the present study. Without establishing the internal and external validity of the present study, by means of the methodological inferences, further consideration of the psychometric inferences would not be justified. Assessment of the methodological inferences is provided within the context of the present study (i.e., exploratory methodological research in psychometrics) and includes a summary of the methodological inferences as well as enumeration of the strengths and weaknesses of the methodological inferences. Subsequent assessment of the psychometric inferences is provided within the context of the present study (i.e., exploratory methodological research in psychometrics) and consistent with the purpose of the present study (i.e., to assess the supplemental ability of thinking-aloud data in the psychometric evaluation of the validity and reliability of aptitude examination items. Assessment of the psychometric inferences includes a summary of the psychometric inferences as well as enumeration of the strengths and weaknesses of the psychometric inferences.

Methodological Inferences

The internal validity of the present study was assessed by means of within-method triangulation, with respect to the

three principal components of the present study: the subjects as the data sources, the nonschedule standardized interview as the means of data collection, and the investigator as the content analyst. The criteria against which the internal validity was assessed included whether the manner in which subjects anticipated responding to the items corresponded to the manner in which subjects actually responded to the items, as well as whether the responses of the subjects to the various aspects of the nonschedule standardized interview revealed interindividual differences in content and comprehensiveness, as described in the Content Analysis section of the Methodology chapter. For all 25 of the items utilized in the present study, the internal validity of the data source, data collection, and data analysis components were considered supported, as was presented in the respective Within-Method Triangulation sections of the Summary of Methodological Inference tables in this chapter (see even numbered tables).

Although the internal validity of the present study was supported by means of within-method triangulation and with respect to the criteria delineated above, the internal validity of the present study would have been further supported had two additional aspects been capable of being considered or addressed. The internal validity of the present study would have been further supported had additional criteria by which to assess the internal validity been identified. Such

additional criteria would have served to enhance the generalizations concerning the internal validity of the present study. The internal validity of the present study would likewise have been further supported had another investigator been a content analyst of the transcripts of the subjects' responses. The methodological inferences derived by another independent investigator content analyst would have served to enhance the generalizations concerning the internal validity of the present study.

The external validity of the present study was assessed by means of between-method triangulation, with respect to the same data source, data collection, and data analysis components utilized in assessing the internal validity. The criteria against which the external validity was assessed included whether the strategies utilized by subjects in responding to the items paralleled those described in the literature, whether the responses of the subjects to the items paralleled task analyses relevant to the items and identified in the literature, and/or whether the psychometric inferences derived for the items paralleled those described in the literature. From one to all three of these criteria were applicable to the 25 items utilized in the present study. For all 25 items, the external validity of the data source, data collection, and data analysis components were considered supported by the criteria applicable, as was presented in the respective Between-Method Triangulation sec-

tions of the Summary of Methodological Inference tables in this chapter.

Although the external validity of the present study was supported by means of between-method triangulation and with respect to the criteria delineated above, the external validity of the present study would have been further supported had two additional aspects been capable of being considered or addressed. The external validity of the present study would have been further supported had all three of the criteria delineated above been identified in the literature for all items and/or had additional criteria by which to assess the external validity been identified. Such additional criteria would have served to enhance the generalizations concerning the external validity of the present study. The external validity of the present study would likewise have been further supported had another investigator been a content analyst of the transcripts of the subjects' responses. The methodological inferences derived by another independent investigator content analyst would have served to enhance the generalizations concerning the external validity of the present study.

Psychometric Inferences

Given that the internal and external validity of the present study, in terms of the data source, data collection, and data analysis components, were considered supported, consideration of the psychometric inferences was seemingly war-

ranted. For the 25 items utilized in the present study, the thinking-aloud responses of the subjects were considered to provide supplemental data to the psychometric data available for each item (i.e., the operational definition of the aptitude purported to be measured by the item) across all three types of psychometric inferences (i.e., content/construct validity, internal consistency/discrimination, alternate form/test-retest reliability).

The psychometric inferences relevant to the content/construct validity of the items suggested that various sources of variance, other than that specified in the operational definition for the item, were determinants in responding to the item. For some items (e.g., GRE/ALR/V/24), in conjunction with the aptitude purported to be measured by the item (i.e., analytical ability/analytical reasoning), sources of variance further included, yet were not restricted to, familiarity with the type of item. For some items, (e.g., KIT/XU3/1/2), in conjunction with the aptitude purported to be measured by the item (i.e., flexibility of use), sources of variance further included, yet were not restricted to, the capacity of memory (i.e., for the restrictions included in the directions for the item, for which item responses had already been listed) as well as experience/acculturation and/or vocabulary.

The psychometric inferences relevant to the internal consistency/discrimination of the items served to corroborate

the content/construct validity inferences and further suggested that, in conjunction with the aptitudes delineated in the Content/Construct Validity sections, sources of variance in item scores included random errors of measurement, in that correct item responses were analogous to the "right answer for the wrong reason" (e.g., visualization, item KIT/VZ3/2/8; inductive reasoning, item KIT/I3/1/7). The psychometric inferences relevant to the internal consistency/discrimination of the items further suggested than manifested interindividual differences in aptitudes and/or strategies did not necessarily correspond to interindividual differences in item scores (e.g., logical reasoning, items KIT/RL4/1/4, KIT/RL4/1/4a; integrative processes, item KIT/IP1/1/9; expressional fluency, item KIT/FE1/2/18).

The psychometric inferences relevant to alternate form/test-retest reliability suggested that presumably parallel items were not necessarily parallel in terms of sources of variance. For some presumably parallel items, the content/construct validity inferences were not parallel between and/or among the items (e.g., logical reasoning, items KIT/RL1/1/2, KIT/RL3/1/9, KIT/RL4/1/4). For some presumably parallel items, the internal consistency/discrimination inferences were not parallel between and/or among the items (e.g., verbal ability/sentence completion, items GRE/VSC/I/3, GRE/VSC/I/3a), either within an alternate form or a test-retest context. For certain items, parallel items would vary

in difficulty dependent upon source of variance not explic-in the operational definition of the item (e.g., logical reasoning, item GRE/ALR/V/25, with variation in the difficulty of parallel items dependent upon reading comprehension and the "concrete" versus "abstract" style in which the paragraph, serving as the basis for the item, was written).

Although the thinking-aloud responses of the subjects were considered to provide supplemental data to the psychometric data available for each item, the psychometric inferences derived in the present study would have been further enhanced had three additional aspects been capable of being considered or addressed. First, the psychometric inferences would have been enhanced had the methodology for the present study not inherently restricted the sample sizes of both subjects and items. Second, the psychometric inferences would have been enhanced had the investigator possessed more expertise in the "factor analytic" interpretation of the aptitudes purported to be measured by the items. Third, had psychometric data other than the operational definitions for the items been available or obtainable (e.g., item analysis indices) for a sample of subjects comparable to the subjects utilized in the present study, the supplemental ability of the psychometric inferences to the other psychometric data would have been enhanced.

Thus, within the context of the present study (i.e., exploratory methodological research in psychometrics), the

results of the present study suggested that the thinking-aloud responses of subjects, as a supplement to the psychometric assessment of aptitude item validity and reliability, constituted both an internally and externally valid methodology. With respect to the purpose of the present study (i.e., to assess the supplemental ability of thinking-aloud data in the psychometric evaluation of aptitude item validity and reliability), the results of the present study suggested that thinking-aloud data possess such a capability when applied to relatively random, though restricted, samples of both items and subjects.

CHAPTER V

SUMMARY AND CONCLUSIONS

The results of the present study seemingly support the premise that thinking-aloud data have the ability to supplement the psychometric assessment of aptitude examination item validity and reliability. However, the results of the present study further suggested that the utility of thinking-aloud data, as a supplement to the traditional psychometric assessment of aptitude measures (i.e., both items and examinations), must be considered in terms of both potential and practical utility.

In terms of potential utility, the supplemental ability of thinking-aloud data to the psychometric assessment of item validity and reliability derives from the assumptions underlying qualitative analysis of item validity and reliability, in contrast to the assumptions underlying quantitative analysis of item validity and reliability (i.e., thinking-aloud data in contrast to psychometric data). Qualitative analysis of item validity and reliability allows assessment of relevant sources of variance in aptitude measures at the level of subjects, items, and/or administrations. By virtue of not being referenced to a given theoretical or mathematical model, qualitative psychometric

analysis allows for detection of multiple sources of variance in aptitude measures (e.g., aptitudes, strategies, random errors of measurement) within and between subjects, items, and/or administrations. Qualitative psychometric analysis further allows detection of the manner in which such multiple sources of variance affect the outcome measures (i.e., item responses, item scores). That is, qualitative psychometric analysis allows detection of whether the multiple sources of variance are linearly or nonlinearly related, are continuous or discontinuous, are interactive or confounding. Furthermore, qualitative psychometric analysis enables assessment of the validity and reliability of items with or without item score variance among subjects. Thus, qualitative psychometric analysis is in contrast to quantitative psychometric analysis, which considers aptitude measures as univariate measures; with sources of variance partitioned into "true" and "error" variance, attributable to interindividual differences in a given aptitude and to random errors of measurement, respectively; with prerequisite score variance; and with interpretation within the context of the mathematical model of linear regression.

With respect to the practical utility of qualitative psychometric assessment of item validity and reliability, as a supplement to quantitative psychometric assessment of item validity and reliability, there is no readily apparent reason to anticipate that the methodology utilized in the

present study would not be equally applicable to other instances or circumstances. Comparable supplemental inferences could seemingly be derived for other so-called objective or "pencil-and-paper" measures, including, but not restricted to "classroom" achievement measures, professional certifying/credentialing examinations, even measures such as personality inventories. Supplemental inferences could seemingly likewise be derived for so-called psychomotor or "practical", "hands-on" measures. Inferences as to why given items perform "well" or "poorly" (i.e., within the context of item analysis indices), why given items manifest "bias" (i.e., within the context of "culture-free" aptitude measures), and why given subjects perform "well" or "poorly" (i e., within the context of diagnosis and/or remediation) could seemingly be derived by means of the methodology utilized in the present study and would correspondingly provide supplemental data relevant to issues such as these.

With respect to the practical utility of qualitative psychometric assessment of item validity and reliability, however, certain limitations, or perhaps more appropriately termed disadvantages, were suggested by the results of the present study. Disadvantages would undoubtedly consist of the amount of time required for the collection and analysis of the thinking-aloud data, as well as the inherent "small sample" restriction for both subjects and items. Given such disadvantages, even though the thinking-aloud data was con-

sidered to have supplemented the operational definitions for the items utilized in the present study to some "significant" degree, an index analogous to a "cost/benefit ratio" is indeterminate. On a routine, comprehensive, exhaustive basis, qualitative psychometric assessment of item validity and reliability would be precluded, given the prohibitive amount of time required, particularly for measures (i.e., aptitude or other types) that are primarily intended for "one-time administrations" (e.g., "classroom" achievement examinations; standardized examinations administered periodically and as revised "editions", for purposes of examination security). However, qualitative psychometric assessment of the validity and reliability of even such measures could be accomplished by means of a purposive or random sample of both subjects and items (e.g., a matrix sampling strategy, such as was utilized in the present study), in order to "screen" or "pretest" measures or in order to sensitize item and/or examination authors and publishers to certain "generic" concepts, which would be applicable or transferable to other items, examinations, or circumstances.

Nonetheless, the results of the present study underscore the fact that the relatively exclusive reliance on quantitative or psychometric assessment of the validity and reliability of aptitude measures provides an incomplete and/or inadequate assessment. The results of the present study suggest that thinking-aloud data serve to supplement the

quantitative or psychometric assessment of validity and reliability of aptitude measures, at the level of items. Thus, within the context of exploratory methodological research in psychometrics, the results of the present study indicate that thinking-aloud data and qualitative psychometric analysis of item validity and reliability exhibit potential utility as a supplement to the quantitative psychometric assessment of item validity and reliability, however, may be limited in terms of practical utility, at least on a routine, comprehensive basis.

The results of the present study further underscore that the descriptions, discussions, and other information (i.e., both nonquantitative and quantitative) provided by the publishers of aptitude measures is incomplete, as presented in the manuals or bulletins which accompany such measures. At least for the two source examinations utilized in the present study, further descriptions, discussions, and information (i.e., both nonquantitative and quantitative) is unavailable from the publishers and not provided in the reference citations compiled by the publishers and appearing in the manuals or bulletins accompanying such measures. The unavailability of further information relevant to the validity and reliability of the aptitude measures exists in spite of statements such as the following:

> ... use of ETS-developed [Educational Testing Service] tests places on the publisher more than ever the responsibility for offering adequate research to support the

> recommended uses of these measures (Ektrom et al., 1976b, p. 6).

Thus, regardless of whether the responsibility for providing more extensive data relevant to the utilization and interpretation of aptitude measures is self-imposed by examination publishers, imposed by professional mandates/guidelines (e.g., Standards for Educational and Psychological Tests and Manuals), and/or imposed by legislation (e.g., "test disclosure laws"), more extensive data relevant to the utilization and interpretation of aptitude measures is presently not available to investigators, and any responsibility for documenting the validity and reliability of aptitude measures seemingly resides, by default, with investigators.

A number of studies identified in the literature had utilized aptitude measures similar or identical to the two source examinations utilized in the present study (e.g., Kropp and Stoker, 1966; Poole, 1971; Sternberg, 1977; French, 1957, 1965; Green et al., 1953). However, none of these studies had included an "assessment" of the validity and the reliability of the aptitude measures utilized. Given the results of the present study, one wonders to what extent the results of those studies might have been interpreted differently had supplemental data, such as that derived in the present study, been available.

REFERENCES

Bloom, B.S. (Ed.). (1956). Taxonomy of educational objectives, handbook I: Cognitive domain. New York: David McKay.

Bloom, B.S., & Broder, L.J. (1950). Problem-solving processes of college students. Chicago: University of Chicago Press.

Bower, G.H., & Hilgard, E.R. (1981). Theories of learning (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Brody, E.B., & Brody, N. (1976). Intelligence: Nature, determinants, and consequences. New York: Academic Press.

Butcher, H.J. (1970). Human intelligence: Its nature and assessment. London: Methuen.

Carroll, J.B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect". In L.B. Resnick (Ed.), The nature of intelligence. Hillsdale, NJ: Lawrence Erlbaum.

Dailey, J.T. (1959). The Graduate Record Examinations Aptitude Test. In O.K. Buros (Ed.), The Fifth Mental Measurements Yearbook. Highland Park, NJ: Gryphon Press.

Denzin, N.K. (1978). The research act: A theoretical introduction to sociological methods. New York: McGraw-Hill.

Detterman, D.K. (1979). A job half done: The road to intelligence testing in the year 2000. In R.J. Sternberg & D.K. Detterman (Eds.), Human intelligence - Perspectives on its theory and measurement. Norwood, NJ: Ablex.

Downie, N.W., & Heath, R.W. (1970). Basic statistical methods (4th ed.). New York: Harper & Row.

Edwards, A.L. (1976). An introduction to linear regression and correlation. San Francisco: W.H. Freeman.

Ekstrom, R.B., French, J.W., & Harman, H.H. (1976a). Kit of Factor-Referenced Cognitive Tests. Princeton, NJ: Educational Testing Service.

Ekstrom, R.B., French, J.W., & Harman, H.H. (1976b). Manual for Kit of Factor-Referenced Cognitive Tests. Princeton, NJ: Educational Testing Service.

Ekstrom, R.B., French, J.W., & Harman, H.H. (1979). Cognitive factors: Their identification and replication. Multivariate Behavioral Research Monographs, 79-2.

Educational Testing Service. (1982). GRE 1982-83 information bulletin. Princeton, NJ: Author.

Fareed, A.A. (1971). Interpretative responses in reading history and biology: An exploratory study. Reading Research Quarterly, 6, 493-532.

Fleishman, E.A. (1975). Toward a taxonomy of human performance. American Psychologist, 30, 1127-1149.

Frederiksen, C.H. (1969). Abilities, transfer, and information retrieval in verbal learning. Multivariate Behavioral Research Monographs, 69-2.

French, J.W. (1957). The factorial invariance of pure-factor tests. Journal of Educational Psychology, 46, 93-109.

French, J.W. (1965). The relationship of problem-solving styles to the factor composition of tests. Educational and Psychological Measurement, 25, 9-28.

Green, R.F., Guilford, J.P., Christensen, P.R., & Comrey, A.L. (1953). A factor-analytic study of reasoning abilities. Psychometrika, 18, 135-160.

Guilford, J.P. (1967). The nature of human intelligence. New York: McGraw-Hill.

Hays, W.L. (1973). Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart and Winston.

Hunt, E., & MacLeod, C.M. (1979). The sentence-verification paradigm: A case study of two conflicting approaches to individual differences. In R.J. Sternberg & D.K. Detterman (Eds.), Human intelligence - Perspectives on its theory and measurement. Norwood, NJ: Ablex.

Humphreys, L.G. (1974). The misleading distinction between aptitude and achievement tests. In D.G. Green (Ed.), The aptitude-achievement distinction. Monterey, CA: CTB/McGraw-Hill.

Humphreys, L.G. (1976). A factor model for research on intelligence and problem solving. In L.B. Resnick (Ed.), The nature of intelligence. Hillsdale, NJ: Lawrence Erlbaum.

Huttenlocher, J. (1976). Language and intelligence. In L.B. Resnick (Ed.), The nature of intelligence. Hillsdale, NJ: Lawrence Erlbaum.

Kaufman, A.S. (1981). The WISC-R and learning diaabilities assessment: State of the art. Journal of Learning Disabilities, 14, 520-526.

Kavale, K., & Schreiner, R. (1979). The reading processes of above average and average readers: A comparison of the use of reasoning strategies in responding to standardized comprehension measures. Reading Research Quarterly, 15, 102-128.

Kerlinger, F.N. (1973). Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart and Winston.

Krippendorff, K. (1980). Content analysis: An introduction to its methodology. Beverly Hills, CA: Sage.

Kropp, R.P., & Stoker, H.W. (1966). The construction and validation of tests of the cognitive processes as described in the taxonomy of educational objectives. Florida State University, Institute of Human Learning and Department of Educational Research and Testing. (ERIC Document Reproduction Service No. ED 010 044).

Kropp, R.P., Stoker, H.W., & Bashaw, W.L. (1966). The validation of the taxonomy of educational objectives. The Journal of Experimental Education, 34, 69-76.

Lerner, R.M. (1976). Concepts and theories of human development. Reading, MA: Addison-Wesley.

Lieberman, D.A. (1979). Behaviorism and the mind: A (limited) call for a return to introspection. American Psychologist, 34, 319-333.

McGrath, E. (1982, December). The "fuzzies" meet the "techs." Time, p. 61.

Morrison, E.J. (1960). On test variance and the dimensions of the measurement situation. Educational and Psychological Measurement, 20, 231-250.

Mukherjee, B.N. (1975). The factorial structure of Wechsler's pre-school and primary scale of intelligence at successive age levels. British Journal of Educational Psychology, 45, 214-226.

Naglieri, J.A., Kaufman, A.S., & Harrison, P.L. (1981). Factor structure of the McCarthy scales for school-age children with low GCIs. Journal of School Psychology, 19, 226-232.

Newell, A., & Simon, H.A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. (1975). SPSS: Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill.

Nunnally, J.C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Olshavsky, J.E. (1976-1977). Reading as problem solving: An investigation of strategies. Reading Research Quarterly, 12, 654-674.

Patton, M.Q. (1980). Qualitative evaluation methods. Beverly Hills, CA: sage.

Pellegrino, J.W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. In R.J. Sternberg & D.K. Detterman (Eds.), Human intelligence - Perspectives on its theory and measurement. Norwood, NJ: Ablex.

Poole, R.L. (1971). Characteristics of the taxonomy of educational objectives: Cognitive domain. Psychology in the Schools, 8, 379-383.

Popham, W.J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

Resnick, L.B. (1976). Introduction: Changing conceptions of intelligence. In L.B. Resnick (Ed.), The nature of intelligence. Hillsdale, NJ: Lawrence Erlbaum.

Reynolds, C.R., & Jensen, A.R. (1983). WISC-R subscale patterns of abilities of blacks and whites matched of full scale IQ. Journal of Educational Psychology, 75, 207-214.

Seddon, G.M. (1978). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. Review of Educational Research, 48, 303-323.

Snow, R.E. (1979). Theory and method for research on aptitude processes. In R.J. Sternberg & D.K. Detterman (Eds.), Human intelligence - Perspectives on its theory and measurement. Norwood, NJ: Ablex.

Standards for educational and psychological tests and manuals. (1966). Washington, D.C.: American Psychological Association.

Sternberg, R.J. (1974). Barron's how to prepare for the Miller Analogies Test (MAT). Woodbury, NY: Barron's Educational Series.

Sternberg, R.J. (1977). Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Lawrence Erlbaum.

Swinton, S.S., & Powers, D.E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. Journal of Educational Psychology, 75, 404-415.

Thorndike, R.L., & Hagen, E.P. (1977). Measurement and evaluation in psychology and education (4th ed.). New York: John Wiley.

Tyler, L.E. (1979). The intelligence we test - An evolving concept. In L.B. Resnick (Ed.), The nature of intelligence. Hillsdale, NJ: Lawrence Erlbaum.

Webster's ninth new collegiate dictionary. (1983). Springfield, MA: Merriam-Webster.

APPROVAL SHEET

The dissertation submitted by Ann Reed Gaines has been read and approved by the following committee:

Dr. Jack Kavanagh, Director
Associate Professor, Foundations of Education and
Associate Dean, School of Education, Loyola

Dr. Judy Irwin
Assistant Professor, Curriculum and Instruction, Loyola

Dr. Steven Miller
Professor, Foundations of Education and
Chairman, Foundations of Education, Loyola

Dr. Ronald Morgan
Associate Professor, Foundations of Education, Loyola

The final copies have been examined by the director of the dissertation and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the dissertation is now given final approval by the Committee with reference to content and form.

The dissertation is therefore accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

4/19/84
Date

Jack A. Kavanagh
Director's Signature