DOCUMENT RESUME

ED 386 461 TM 023 797

AUTHOR	Engelhard, George, Jr.; And Others
TITLE	Constructing Rater and Writing Task Banks for the Assessment of Written Composition.
PUB DATE	31 Oct 94
NOTE	16p.; Earlier version of a paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Data Collection; *Educational Assessment; *Interrater Reliability; *Item Banks; Item Response Theory; Networks; *Test Construction; Writing (Composition); *Writing Evaluation
IDENTIFIERS	Calibration; FACETS Model; *Large Scale Programs; *Rasch Model; Writing Prompts

ABSTRACT

A set of procedures is described for constructing an assessment network composed of a connected system of rater and writing task banks within the context of large-scale assessments of written composition. The calibration of the assessment tasks and the measurement of individuals are viewed as separate, although complementary, activities. The writing task bank is a calibrated set of prompts with content and measurement characteristics that have been systematically examined and cataloged. A rater bank is a calibrated set of judges whose measurement characteristics have also been systematically examined and cataloged. In large scale assessment, these ideas are extended to networks, with the network being a calibrated measurement system of rater and task banks. The first section of the paper describes an extended version of the Rasch model, the FACETS model, that can be used to construct a consistent and coherent assessment network. The next section presents illustrative data collection designs that may be used to calibrate an assessment network. Three tables and two figures illustrate the discussion. (Contains 22 references.) (SLD)

CONSTRUCTING RATER AND WRITING TASK BANKS FOR THE ASSESSMENT OF WRITTEN COMPOSITION

George Engelhard, Jr.

Emory University

Belita Gordon and David Curtin

University of Georgia

Address: Professor George Engelhard, Jr.
Division of Educational Studies
210 Fishburne Building
Emory University
Atlanta, GA 30322

(404) 727-0607
(404) 525-1115 h

Running head: Rater and writing task banks

October 31. 1994

[RaterTaskBank.AERA]

An earlier version of this paper was presented at the annual meeting of the American Educational Research Association (April 1994) in New Orleans.

CONSTRUCTING RATER AND WRITING TASK BANKS FOR THE ASSESSMENT OF WRITTEN COMPOSITION

The purpose of this paper is to describe a set of procedures for constructing an assessment network composed of a connected system of rater and writing task banks within the context of large-scale assessments of written composition. The ideas presented here grew out of our work on the development of a large-scale assessment program for the measurement of writing competence within the context of a high school graduation test. One of the major goals of this work has been to develop a calibrated set of raters and writing tasks that can be used for the objective measurement of writing competence. In order to accomplish this goal, we have focused on meeting the requirements of objective measurement within the framework of the Rasch model (Engelhard, 1992; 1994). We have found it useful to view the calibration of the assessment tasks and the measurement of individuals as separate, although complementary, activities. This approach is congruent with accepted measurement practices; typically, measurement practitioners first calibrate their instruments, and then administer these instruments along with appropriate checks on whether or not each examinee is being assessed objectively and fairly. An assessment network depends in a fundamental way on the measurement model selected, as well as the data collection design used to calibrate the facets of the assessment network.

In a series of papers, Choppin (1968,1978, 1982) described how item banks can be used to contribute to the improvement of measurement. Choppin defines an item bank as follows:

> "The term 'item bank' should be understood to mean a collection of test items organised and catalogued in a similar way to books in a library. This organising and cataloguing takes account of the content of the test item and also its measurement characteristics (such as difficulty, reliability, validity, etc.). Such items can be readily grouped into tests which will then be properly defined and calibrated measuring instruments" (Choppin, 1978, p. 1).

Based on this definition of item banks, a writing task bank can be defined as a calibrated set of prompts whose content and measurement characteristics have been systematically examined and cataloged. In a similar fashion, a rater bank can be defined as a calibrated set of judges whose measurement characteristics have been systematically examined and cataloged. In large-scale performance assessments, it useful to extend this idea to include networks (Engelhard & Osberg, 1983) with an assessment network defined as a calibrated measurement system composed of rater and writing task banks. In the language of ANOVA, the crossing of the rater and writing task banks yields an assessment network that is composed of a variety of assessment components; each assessment component yields an assessment opportunity for an examinee to obtain an observed rating or score. This paper builds upon and extends the idea of item banks to include both writing task and rater banks, as well as the construction of an assessment network composed of a coherent set of banks. In terms of the classification system for linking procedures proposed by Mislevy (1992), the procedures described in this paper reflect calibration more closely than equating.

In the first section of this paper, an extended version of the Rasch model is described that can be used to construct a consistent and coherent assessment network. In the next section, illustrative data collection designs that may be used to calibrate an assessment network are described.

A FACETS MODEL FOR WRITING ASSESSMENT

The general model for the assessment of written composition that guides this paper is presented in Figure 1. Ideally, writing competence should be the major variable affecting the observed rating.

Insert Figure 1 about here

In practice, when the measurement of writing competence is based directly on student compositions, there are a variety of factors, such as rater and writing task characteristics, that may be viewed as intervening variables. The assessment process should minimize, as much as possible, the effects of these intervening variables on the estimates of writing competence. The situation becomes even more complex when different students are rated by different raters who may vary in severity, and also when different students respond to different writing tasks that may vary in difficulty. The development of rater and writing task banks provides the opportunity to statistically adjust for these differences that may appear when students are not rated by all of the raters on all of the writing tasks, and to obtain fairer and more objective estimates of student competence in writing.

The procedures described here for constructing an assessment network composed of rater and writing task banks are based on a multifaceted version of the Rasch measurement (FACETS) model for ordered response categories developed by Linacre (1989). The FACETS model is an extended version of the Rasch measurement model (Andrich, 1988; Rasch, 1980; Wright & Masters, 1982). In essence, the FACETS model is an additive linear model based on a logistic transformation of the observed ratings to a logit scale. Using the terminology of regression analysis, the dependent variable is the logistic transformation of ratios of successive category probabilities (log odds), and the independent variables are the facets. For example, if writing competence was measured with several writing tasks with the compositions rated as pass or fail, then an appropriate Rasch model for this dichotomous data can be written as follows:

$$\ln [P_{ni1}/P_{ni0}] = \beta_n - \delta_i$$

where

P_{ni1} = probability of student n passing (x=1) on writing task i

P_{ni0} = probability of student n failing (x=0) on writing task i

β_n = Writing competence of student n

δ_i = Difficulty of writing task i.

This model has two facets -- student competence and writing task difficulty. This form of the model can be easily extended to deal with rating scale data and multiple facets. The three-facet model (student competence, writing task difficulty, and judge severity) with four rating categories (0, . . ., 3) used in this paper can be written as follows:

$$\ln[P_{nijk}/P_{nijk-1}] = \beta_n - \delta_i - \lambda_j - \kappa_k \qquad (1)$$

where

P_{nijk} = probability of student n being rated k on writing task i by rater j

P_{nijk-1} = probability of student n being rated k-1 on writing task i by rater j

β_n = Writing competence of student n

δ_i = Difficulty of writing task i

λ_j = Severity of rater j

κ_k = Difficulty of rating Step k relative to Step k-1.

The rating scale parameter, κ_k, which reflects the structure of the four-category rating scale used in this paper are not labelled as a facet in the model.

The FACETS model is a unidimensional model with a single student competence facet, and a collection of other assessment facets, such as writing task and raters. The crossing of these assessment facets defines a set of assessment components that yield multiple ratings for each student. For example, if students responded to two writing tasks and the compositions were rated by three raters, then the assessment network would consist of six assessment components with six observed ratings for each student. The FACETS model is appropriate if the intent of the assessment developers is to sum the ratings from the assessment components in order to produce a total score. As with other Rasch measurement models, the basic assumption of the FACETS model is "that the set of people to be measured, and the set of tasks (items) used to measure them, can each be uniquely ordered in terms respectively of their competence and difficulty" (Choppin, 1987, p. 111). If the data fit the model and this unique ordering is realized, then a variety of desirable measurement characteristics can be attained. Some of these measurement characteristics are (1) separability of parameters with sufficient statistics for estimating these parameters, (2) invariant estimates of student competence, rater severity and writing task difficulty (this reflects the property of "specific objectivity in Rasch's terminology), and (3) equal-interval scales for the measures. Another way to think about the construction of an assessment network with the FACETS model is to view it as an "equating model" with the raters and writing tasks viewed as analogous to test forms that may vary in difficulty; if different students are rated by different raters on different writing tasks, then it may be necessary to "equate" or statistically adjust for differences in rater severity and writing task difficulty.

Based on the FACETS model presented in Equation 1, the probability of student n with competence β_n obtaining a rating of x (x = 0, 1, . . . , m) on writing task δ_i from rater λ_j with category step difficulty of τ_s is given as

$$\pi_{nij0} = \frac{1}{1 + \sum_{k=1}^{m} \exp\left[k\left(\beta_n - \delta_i - \lambda_j\right) - \sum_{s=1}^{k} \tau_s \right]}$$

for x = 0, and

$$\pi_{nijx} = \frac{\exp\left[x\left(\beta_n - \delta_i - \lambda_j\right) - \sum_{s=1}^{x} \tau_s \right]}{1 + \sum_{k=1}^{m} \exp\left[k\left(\beta_n - \delta_i - \lambda_j\right) - \sum_{s=1}^{k} \tau_s \right]}$$

for x = 1, . . . , m.

Linacre (1989) provides a detailed description of the FACETS model, as well as procedures for estimating the parameters of the model. The fit of rating scale data to the FACETS model can be examined in various ways; Wright and Masters (1982) and Wright and Stone (1979) should be consulted for detailed descriptions of the standardized residuals, the INFIT and OUTFIT statistics, and the reliability of separation index.

DESCRIPTION OF THE DESIGNS

There are a variety of data collection designs that can be used to calibrate raters and writing-tasks. In this section, a set of representative designs are described that can be used to illustrate many of the data collection issues that need to be considered in the construction of rater and writing task banks. A complete cataloging of all designs is beyond the scope of this paper. As much as possible, an attempt has been made to construct a bridge between the widely accepted language used with equating traditional multiple-choice tests with several forms (Andrich, 1988; Petersen, Kolen, & Hoover, 1989), and the language used with calibrating IRT models, such as the Rasch model (Hambleton, Swaminathan, & Rogers, 1991; Linacre, 1989; Wright & Stone, 1979). The measurement situation used to illustrate the designs is based on two writing tasks, three raters, and ten examinees; the extensions of these designs and basic principles to assessment networks with more than three facets are straightforward. Of course, operational designs for calibrating writing tasks and raters would be based on many more examinees and usually more raters. In essence, examinees can be viewed as replications within each cell of the design, and increasing the number of examinees within a cell would result in a concomitant decrease in the standard error for any estimates that included that cell. There are three general categories of designs that can be used for linking together assessment components into a consistent and coherent network. These categories are complete, incomplete, and non-linked assessment networks.

Before describing the designs, it is useful to define more a clearly a few of the terms. ***Facets*** are defined as the separate dimensions that are used in the assessment network. Within the language of the analysis of variance, facets are similar to factors. Facets are composed of individual ***elements*** that vary in difficulty, and the difficulty of an element defines its location on the latent variable that the assessment network is designed to measure. For example, each writing task is an element within the writing-task facet, and each rater is an element within the rater facet. It should also be noted that the examinee is considered a facet in this model, while in Generalizability Theory examinees are not considered a "facet" (Shavelson and Webb, 1991). When rater and writing task facets are crossed, then the cells within the design are called ***assessment components***; each assessment component yields an assessment opportunity for the examinee to obtain a observed rating that depends on the difficulty of the elements from each facet that combine to define that cell. The assessment components obtained from a crossing of several facets combine to define an overall ***assessment network***.

Complete assessment networks consist of completely crossed designs with examinees having observed scores on all of the assessment components. Examples of these designs are shown in Table 1. These completely crossed designs are the simplest data collection designs. Since all of the

Insert Table 1 about here

examinees have observed scores from all of the assessment components. the writing competence of the examinee is not confounded with the calibration of the assessment components. The connectedness of complete assessment networks can be presented graphically as shown in the first column of Figure 2.

Insert Figure 2 about here

The circles represent the assessment components, and the lines indicate that examinee data is available that provides for the direct estimation of a link between all of the assessment components included in the overall assessment network. In practice, it would be desirable to randomize the order of presentation of the writing tasks to the examinees; this would help to minimize the effects of extraneous factors, such as learning, fatigue and practice. Context effects may also influence the rating behavior of the raters, and the order of the presentation of the compositions to the raters should also be randomized. The number of assessment components for the Two-Facet Designs (task x examinee and rater x examinee) match the number of elements (tasks or raters) in the design. For the Three-Facet Design, the number of assessment components reflects the product of the number of raters times the number of writing tasks (3 x 2 = 6). These designs for constructing complete assessment networks are essentially generalizations of the Single-Group and Counterbalanced Random Groups Designs described by Petersen, Kolen, and Hoover (1989).

Incomplete assessment networks consist of designs in which examinees do not have scores on all of the assessment components, and systematic links have to be created in order to yield a connected network of assessment components. When developing a calibrated assessment network, there are a variety of practical considerations that rule out the construction of complete assessment networks. Carefully designed incomplete assessment networks can be used to obtain reliable and valid links both within and between facets that are less costly in terms of examinee time and rater salaries. Examples of these types of designs are shown in Table 2.

Insert Table 2 about here

For two-facet designs (task x examinee or rater x examinee), it is possible to calibrate each facet through common examinees or through an anchor facet (anchor tasks or anchor raters). The number of assessment components and the number of observed ratings obtained for each examinee are not the same in an incomplete assessment network. The connectedness of incomplete assessment networks can be presented graphically as shown in the second column of Figure 2. For incomplete assessment networks, all of the assessment components are linked together, although there are fewer links. The construction of connected incomplete assessment networks is extremely complex, and there are many choices for acceptable designs. In fact, if it is recognized that the data collection designs used to construct incomplete assessment networks are examples of Balanced Incomplete Block (BIB) and Partially Balanced Incomplete Block (PBIB) designs with block sizes of at least two, then there are a plethora of designs that can be considered (John, 1980; Kirk, 1968). BIB and PBIB designs make it possible to estimate "main effects," but the situation becomes more complicated when bias analyses and differential facet functioning based on interactions among the facets need to be explored. If systematic links are not built into the data collection design, then non-linked assessment networks may result; Weeks and Williams (1964) have described a straightforward procedure for identifying linked assessment networks, and this procedure is used in the FACETS computer program to check for connectedness (Linacre & Wright, 1992). Many of these issues also appear in the literature on paired comparisons (David, 1988). These designs reflect generalizations of the Anchor-Test Designs described in Petersen, Kolen, and Hoover (1989).

Non-linked assessment networks consist of designs in which examinees do not have scores on all of the assessment components, and there are no systematic links among the assessment components. Examples of these types of designs are shown in Table 3.

Insert Table 3 about here

The lack of connectedness in non-linked assessment networks can be presented graphically as shown in the third column of Figure 2. These designs lead to assessment networks that break into two or more disconnected networks of assessment components that depend on the nesting structure of the data collection design. These designs have many weaknesses, and some measurement professionals might even question including these designs or even calling them "networks." The quality of the network depends on how well the "equivalent" groups have been defined. In the language of the analysis of variance, the examinees or other facets of the assessment network are nested within other facets. This nesting makes it impossible to directly calibrate the assessment components, and additional assumptions are required to connect the disconnected assessment components. For example, if the writing tasks are not directly linked, then it is not possible to directly eliminate the potential influences of the particular examinees used to calibrate the assessment network. These designs for constructing non-linked assessment networks are essentially generalizations of the Equivalent Groups Designs described by Petersen, Kolen, and Hoover (1989).

DISCUSSION

Item banks have provided a framework for solving a variety of measurement problems (Wright & Bell, 1984). As the number of direct assessments of writing competence increases, it is likely that rater and writing task banks that combine to form coherent and consistent assessment networks can provide a similar framework for improving measurement practice for this type of performance assessment. Our work on rater and writing task banks has been guided by the view that it is necessary to develop a systematic set of procedures and data collections designs which will provide as much control as possible over the quality of the data collected, as well as meet the requirements of objective measurement within the framework of Rasch measurement. In order to achieve objective and fair measurement, it is necessary to develop invariant calibrations of the facets of the overall assessment network. The calibrations of rater severity and writing task difficulty should be sample-invariant, and therefore not depend upon the particular examinees used to obtain these calibrations.

REFERENCES

Andrich, D. (1988). *Rasch models for measurement.* Newbury Park: Sage Publications.

Choppin, B. (1968). An item bank using sample-free calibration. *Nature, 219,* 870-872.

Choppin, B. (1978). *Item banking and the monitoring of achievement.* (Research in Progress Series No. 1). Slough: National Foundation for Educational Research.

Choppin, B. (1982). The use of latent trait models in the measurement of cognitive abilities and skills. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology* (pp. 41-63). Melbourne: Australian Council for Educational Research.

Choppin, B. (1987). The Rasch model for item analysis. In D.L. McArthur (Ed.), *Alternative approaches to the assessment of achievement* (pp. 99-127). Norwell, MA: Kluwer Academic Publishers.

David, H.A. (1988). *The method of paired comparisons.* London: Charles Griffin & Company Limited.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education,* 5(3), 171-191.

Engelhard, G, & Osberg, D.W. (1983). Constructing a test network with a Rasch measurement model. *Applied Psychological Measurement, 7,* 283-294.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications, Inc.

John, P.W.M. (1980). *Incomplete block designs.* New York: Marcel Dekker, Inc.

Kirk, R.E. (1968). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Brooks/Cole Publishing Company.

Linacre, J. M. (1989). *Many-Faceted Rasch Measurement.* Chicago: MESA Press.

Linacre, J. M., & Wright, B.D. (1992). *A user's guide to FACETS: Rasch measurement computer program.* Chicago: MESA Press.

Mislevy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS, Policy Information Center.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational Measurement*, Third Edition (pp. 221-262). New York: American Council on Education and Macmillan Publishing Company.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications, Inc.

Weeks, D.L., & Williams, D.R. (1964). A note on the determination of connectedness in an n-way cross classification. *Technometrics, 6*(3), 319-324.

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement, 21*(4), 331-345.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.

Table 1

Illustrative Data Collection Designs with Complete Assessment Networks

Assessment Component	Rater	Task	Examinee									
			1	2	3	4	5	6	7	8	9	10
1. Two-Facet Design (task x examinee)												
1		1	√	√	√	√	√	√	√	√	√	√
2		2	√	√	√	√	√	√	√	√	√	√
2. Two-Facet Design (rater x examinee)												
1	1		√	√	√	√	√	√	√	√	√	√
2	2		√	√	√	√	√	√	√	√	√	√
3	3		√	√	√	√	√	√	√	√	√	√
3. Three-Facet Design (rater x task x examinee)												
1	1	1	√	√	√	√	√	√	√	√	√	√
2	2	1	√	√	√	√	√	√	√	√	√	√
3	3	1	√	√	√	√	√	√	√	√	√	√
4	1	2	√	√	√	√	√	√	√	√	√	√
5	2	2	√	√	√	√	√	√	√	√	√	√
6	3	2	√	√	√	√	√	√	√	√	√	√

Note. These designs are essentially generalizations of Single-Group Designs (Petersen, Kolen, & Hoover, 1989). Even though the designs are represented here with 10 examinees, operational designs would require more examinees. A √ indicates that a rating is obtained for the examinee on this assessment component; otherwise a rating is not obtained.

Table 2

Illustrative Data Collection Designs with Incomplete Assessment Networks

Assessment Component	Rater	Task	Examinee 1	2	3	4	5	6	7	8	9	10
1. Two-Facet with Common-Examinee Design (task x examinee)												
1		1	√	√	√	√	√	√	√			
2		2				√	√	√	√	√	√	√
2. Two-Facet with Anchor-Rater Design (rater x examinee)												
1	1		√	√	√	√	√					
2	2		√	√	√	√	√	√	√	√	√	√
3	3							√	√	√	√	√
3. Three-Facet with Anchor-Rater Design (rater x task x examinee)												
1	1	1	√	√	√	√	√					
2	1	2	√	√	√	√	√					
3	2	1	√	√	√	√	√	√	√	√	√	√
4	2	2	√	√	√	√	√	√	√	√	√	√
5	3	1						√	√	√	√	√
6	3	2						√	√	√	√	√

Note. These designs are essentially generalizations of Anchor-Test Designs (Petersen, Kolen, & Hoover,1989). Even though the designs are represented here with 10 examinees, operational designs would require more examinees. A √ indicates that a rating is obtained for the examinee on this assessment component; otherwise a rating is not obtained.

Table 3

Illustrative Data Collection Designs with Non-linked Assessment Networks

Assessment Component	Rater	Task	Examinee 1	2	3	4	5	6	7	8	9	10
1. Two-Facet Design (examinee:task)												
1		1	√	√	√	√	√					
2		2						√	√	√	√	√
2. Two-Facet Design (examinee:rater)												
1	1		√	√	√							
2	2					√	√	√	√			
3	3									√	√	√
3. Three-Facet Design (rater x examinee:task)												
1	1	1	√	√	√	√	√					
2	2	1	√	√	√	√	√					
3	3	1	√	√	√	√	√					
4	1	2						√	√	√	√	√
5	2	2						√	√	√	√	√
6	3	2						√	√	√	√	√

Note. These designs are essentially generalizations of Equivalent-Groups Designs (Petersen, Kolen, & Hoover, 1989). Even though the designs are represented here with 10 examinees, operational designs would require more examinees. A √ indicates that a rating is obtained for the examinee on this assessment component; otherwise a rating is not obtained.

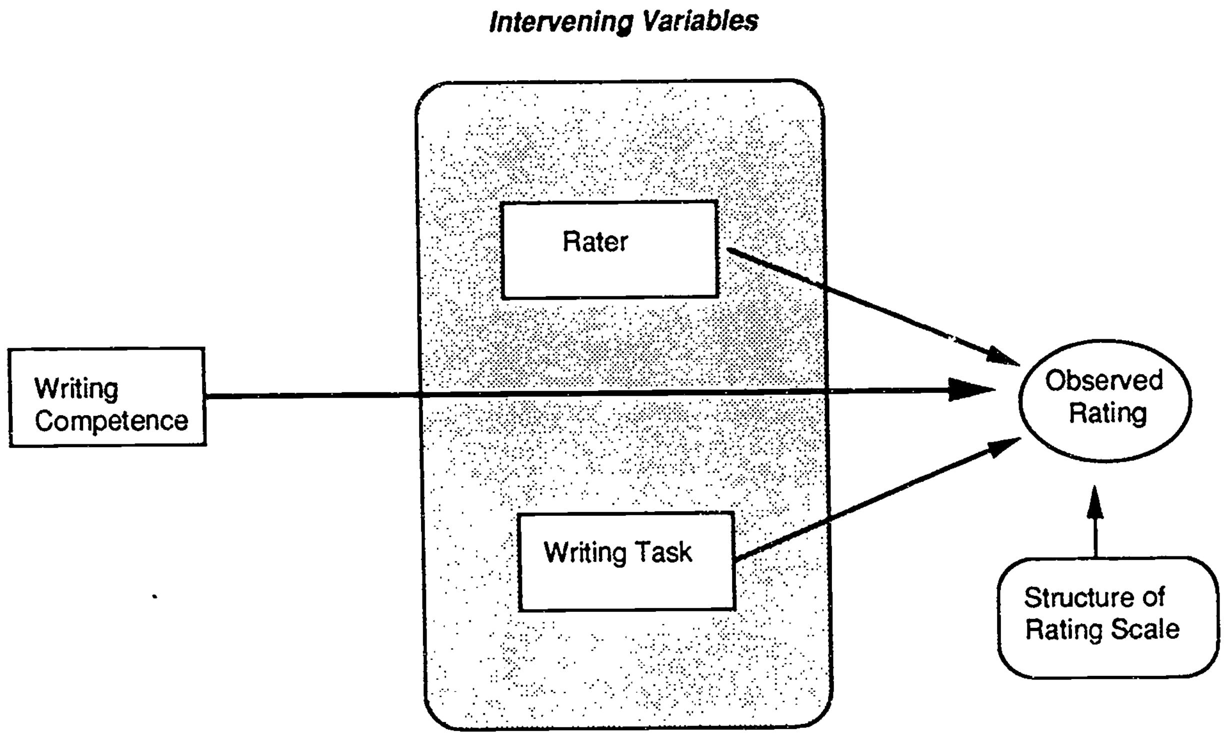

Figure 1

Measurement model for the assessment of writing competence

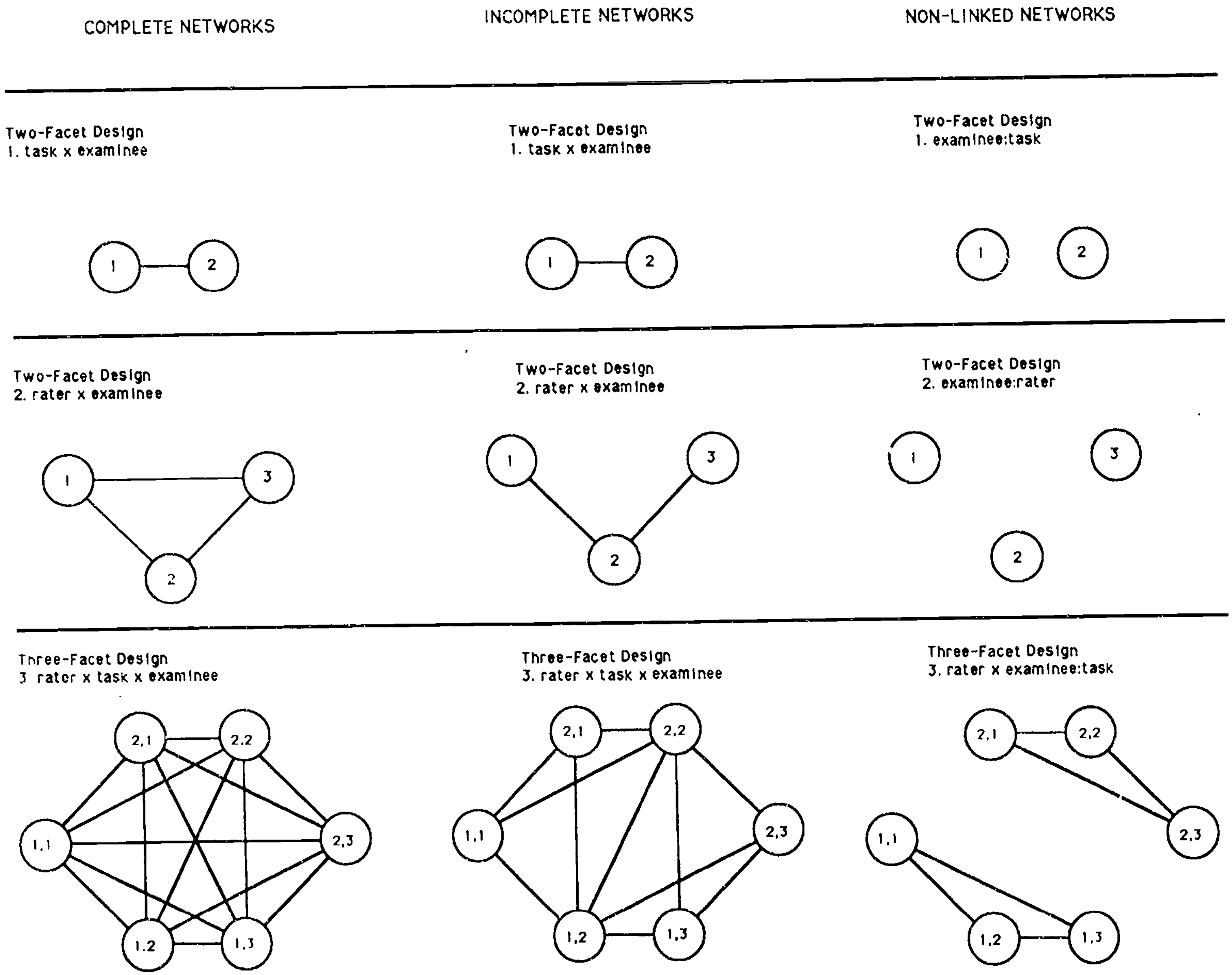

Figure 2 Diagrams of Data Collection Designs