Assessing the Unidimensionality of the NCLEX-RN

Thomas O'Neill and Michelle Reynolds
NCSBN, Inc.

The National Council of State Boards of Nursing, Inc. (NCSBN) is a not-for-profit organization that is composed of the jurisdictional boards of nursing in the United States and US territories. NCSBN's mission is to provide leadership to advance regulatory excellence for public protection. One of the many ways that NCSBN fulfills this mission is by providing its members (boards of nursing) with a defensible method of assessing a candidate's competence. Specifically, NCSBN creates and administers two minimal competency examinations, the National Council Licensure Examination for Registered Nurses® (NCLEX-RN®) and the National Council Licensure Examination for Practical Nurses® (NCLEX-PN®). All boards of nursing that are members of NCSBN use the NCLEX® as part of their licensing process.

Although adaptive tests provide many benefits, they also introduce many challenges. Sparse data is one of the major issues that is both a benefit and a challenge. It is wonderful that candidates may take shorter, less grueling tests because a small subset of the items is all that is needed, but the drawback is that the resulting data matrix is incomplete. In fact, when large item pools are used, the data matrices are quite sparse. With the introduction of the Rasch model (1960) and item response theory (IRT) in general, calibrating items and estimating person ability on incomplete sets of the data is no longer extraordinary. In fact the NCLEX examinations have used Rasch's (1960) model for dichotomous items since 1984 to calibrate test items and measure candidates' ability. Yet one of the requirements of measurement implied by the Rasch model is unidimensionality. Interactions between people and items that result in data that cannot be sufficiently order by a single continuum are multidimensional and degrade the measurement properties of the item calibrations and candidate scores. Therefore, it is important to periodically assess that the interaction of candidates and items is predominantly unidimensional. However, tests of dimensionality typically require complete data or near complete data designs. The methods available to assess multidimensionality in sparse data matrices seem relatively few. The two most popular are analysis of model-data fit and principle components analysis. This paper presents a method for testing the hypothesis of unidimensionality using PCA given a sparse data matrix and an example to illustrate it.

Dimensions

The NCLEX examinations were designed to measure a single construct, nursing ability. Nursing ability could have been conceived of as being composed of several separate constructs (client needs, nursing process, specialty area, etc.), but that approach would require the development of several different scales and passing criteria for each one. Instead, the more general, overarching construct of "nursing ability" which encompasses those more specific areas was selected because it was a more parsimonious model. This paper addresses whether a general construct of nursing ability is warranted given the observed data.

Before investigating whether test data manifests some degree of multidimensionality, it is important to have a clear understanding of what dimensions are and where they come from. A dimension is the imposition of a human organizational schema upon experience in such a way that it is coherent, useful, and represents a single continuum of more or less. Dimensions are not creatures of nature waiting to be discovered; they are abstractions created, selected, and then maintained in the user's mind. People use hierarchies to understand dimensions in terms of more and less, and in the case of dichotomous questions, it is usually a hierarchy of the content of those items.

Although it is obvious that the organizational schema must be coherent and represent a continuum of more to less, it is not as obvious that the dimension is chosen by and sometimes intentionally adjusted by the measurer. When mariners set sail for a destination across the ocean, they must measure both direction and distance to efficiently navigate. However, they choose to use geodesic distance rather than straight-line distance because typically sailors travel over the earth's surface rather than bore holes through the planet. The invention and use of geodesic distance does not invalidate the notion of straight-line distance. Geodesic distance merely represents a better theory of how sailors travel. The selection of a dimension and the organizational schema used to represent that dimension should be governed by the researchers intentions.

METHOD

Unidimensionality is usually assessed by analysis at the form level. For written tests, the analysis is rather straightforward. The items on the test form are tested to see if they are measuring the same thing, often using some form of factor analysis. After a few forms have been analyzed and the unidimensionality is supported, the items are considered to be unidimensional. There are many methods for assessing the dimensionality of complete data sets and the full complement of those procedures will not be described here. When analyzing sparse data matrices as found with CAT data, the solution has typically taken two paths: model-data fit or principle components analysis of residuals. The popularity of these methods has increased largely because of the ease with which they can be performed using standard item calibration software such as Winsteps (Linacre, 2004) and RUMM2020 (Andrich, 2004). The method used in this paper uses PCA of standardized residuals. In addition to being convenient to do, PCA was a logical choice. Although the data set being examined is hoped to be sufficiently unidimensional, it is possible that it has many dimensions. Because the specifics of what those dimensions might be are too numerous to easily manage, a process to identify the largest dimension was needed. PCA meets that requirement.

Procedure

Typically, factor analysis is used to identify or validate subscales within a test or a battery of tests. However, the Rasch model specifies that there is only one latent dimension. Therefore, when factor analysis is used to support a Rasch-based scale, its purpose is not to find a number of factors, but rather to confirm the unidimensionality of the test with a given data set. If there is truly only one latent dimension, then there should not be any latent structure in the residuals and consequently an analysis of the residuals should not yield any noticeable factors. There should only be noise. Because the hypothesis is that only one dimension exists in the data, a factor analytic procedure that extracts the maximum amount of variance onto the first factor is desireable. It is of no consequence what that factor is. The existence of any residual factor degrades the measurement model. Yet because PCA can extract small factors even from randomly generated data, it is necessary to identify how large a factor must be to be considered real. Simulated data can help to answer this question. The procedure basically compares the results of observed data with results from simulated data that is modeled to meet the specifications of the Rasch model and has an identical number of items and examinees.

A person-by-item response matrix is analyzed using Rasch's (1960) model for dichotomous items, calibrating the items and estimating the ability of the candidates. These item and person parameters are then used to create a matrix of expected responses. Observed responses are either 0 or 1, but the expected response can range from 0 to 1, inclusively. Subtracting the expected response from the observed response and dividing the difference by the model standard deviation yields the standardized residual. A principle components analysis is performed on the matrix of standardized residuals to determine the size of the first (largest) factor.

Next, a person-by-item response matrix is simulated. This simulated data set should have the same number of items and people, represent the same items selection rules that were used in observed data set, but have responses that are driven by the requirements of the Rasch model. Specifically, this means that the probability of correctly answering an item is computed based upon the difference between the selected item's difficulty and the selected person's ability. This probability is compared to a randomly generated number selected from a uniform distribution that is bounded by 0 and 1.

PCA of Residuals

One of the dominant ways to detect multidimensionality in a particular data set is to analyze the residuals. Given the ability estimate for each candidate and the difficulty calibration for each item, an expected score for each item can be calibrated. The difference between the observed score (0 or 1) and the expected score (0 through 1) is the residual. A principle components analysis of the residuals can help to detect trends that cannot be attributed to the intended measurement dimension.

However, PCA can extract factors even from random data. Therefore, it is important to have a baseline for what Eigen value must be exceeded before a first factor is identified as being real, rather than as an accident of the data.

Data

Two types of data were analyzed. The first was NCLEX-RN examination results collected from April 1, 2004 to September 30, 2004. During this period there were 89,116 examinees. The second data set was simulated to be comparable to the first data set with regard to the difficulty of the items available, the ability of the candidates testing, and the same rules for item selection and scoring. The simulated dataset was different from the real dataset in that the simulees' responses to the items were model to meet the unidimensional expectations of the Rasch model.

In general, data generated by a computerized adaptive test (CAT) will yield a sparse data matrix. The datasets in this study are no exception. Candidates can receive between 60 and 250 items from a pool of 2,000 items[1]; therefore for each candidate, only 3.0% to 12.5% of the 2,000 possible items will have responses. On average, only 6% of the items had responses, leaving 94% of the real data matrix blank. The simulated data matrix was created to be comparable.

NCLEX-RN

The NCLEX-RN is a variable-length, computerized adaptive test. Each candidate's examination conforms to the current test plan (NCSBN, 2003) and contains 75 to 265 questions. Of these questions, 15 are unscored pretest items. Every time the examinee answers a scoreable question, the computer re-estimates the examinee's ability and subsequently selects a question from the item bank that will both meet the test plan requirements with regard to content and have a level of difficulty that the examinee should find challenging. This provides a test that is well targeted to each examinee. After question 75 is answered, the computer attempts to determine with 95% confidence whether the examinee's true ability is above or below the passing standard. This is accomplished by determining if the candidate's ability estimate is more than 1.67 standard errors away from the passing standard. If it is above, the test stops and the examinee passes. If it is below, then the test stops and the examinee fails. If the computer cannot make a decision with 95% confidence, then it asks another question. This continues until (i) a decision is reached, (ii) the maximum number of items is reached, or (iii) the examinee runs out of time. If an examinee reaches the maximum number of items without a pass-fail decision being made, the 95% certainty requirement is dropped. At the maximum number of items, an examinee's ability estimate is quite precise. Abilities estimates above the passing standard pass. Ability estimates at or below the passing standard fail. If an examinee runs out of time before answering the maximum number of questions, the decision process is more complex. In this case, the examinee's ability estimate on the last item is compared with the passing standard. If it is not above passing, the examinee fails. If it is above passing, then the examinee's ability estimate on the second to last item is compared to the passing standard. If this estimate is also above passing, then the third to last ability estimate is compared to the passing standard. This process continues over the last 60 ability estimates. If the examinee's ability estimate drops to or below the passing standard even once on the last 60 items, the examinee fails[2].

Every operational question in the item bank has undergone repeated review with regard to content and has met all of NCSBN's statistical requirements. The items are calibrated using a one-parameter logistic model, Rasch's (1960/1980) model for dichotomous questions.

Rasch Model

All Rasch models are logistic, latent trait models of probability for monotonically increasing functions. These models are derived not from data but from the structure necessary for measurement. Consequently, the Rasch model is imposed on data. This is quite different from "statistical" approaches in which a model is created to efficiently summarize or reproduce the observed data. The model demands that when two

[1] Of the 2000 items in the operational pool, 27 of them were turned off for formatting or content reasons before the test was ever administered. As a result, there were only responses to 1,973 items in the observed data.

[2] An ability estimate at any point on the test is based upon the responses to all items up to that point. Therefore, it would be incorrect to say that the "last 60 rule" considers only the responses on the last 60 items. It is also important to keep in mind that the maximum item rule and the "last 60 rule" are essentially a second chance for those examinees that were not able to meet the requirement of demonstrating their competence with 95% certainty.

people of different ability encounter an item, the person with the higher ability ALWAYS has the higher probability of answering it correctly. Similarly, when a person encounters two items of different difficulty, the more difficult item ALWAYS has a lower probability of being answered correctly than the easier one. The philosophy behind Rasch's model is that there is a single continuum onto which both items and people are mapped. Because the items represent what the examinee can and cannot do, the ordering and relative spacing of the items articulates the construct. Subsequently, a person's ability estimate is then expressed as the point on that item continuum where the person has a 50-50 chance of correctly answering an item. It is immediately obvious that the invariance of the item hierarchy is crucial.

The dichotomous Rasch model specifies that the probability of a correct response is governed by the difference between the ability of the person, β_v and the difficulty of the item, δ_i. However, the difference (β_v- δ_i) can range from infinity to negative infinity, but the probability of a correct response is limited to the range of zero to one. Converting the probability to a log odds ratio solves the restriction of range problem. Expressed mathematically, the dichotomous Rasch model is specified as:

$$\Pr\{(x_{ni1}) \mid \beta_n, \delta_i\} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} \qquad (1)$$

where

Pr_{ni1} is the probability of a correct response
(Pr_{ni0} would be the probability of an incorrect response),
β_n is the ability of person *n*,
δ_i is the difficulty of item *i*, and
e is the base of the natural log function.

Because the model requires that the relative difficulty of the items remain stable, responses by individuals or groups of individuals that grossly violate that notion can be detected statistically through a variety of procedures such as person misfit, item misfit, parameter drift, differential item functioning, and others.

Rasch's model separates the person and item parameters, yet expresses them on the same scale. As a result, the same person ability estimate should be derived regardless of the particular items administered. This is true regardless of the overall difficulty of the test. Similarly, item difficulty calibrations should be the same regardless of the particular people who answered the question. This is true even when items are calibrated on groups of people with noticeably different mean abilities. Notice that the requirements of sampling theory (random assignment to create equal groups, normal distributions, interval scale observations, etc.) are not required for the Rasch model. When the responses fit the Rasch model, interval measurement is achieved and a stable construct is articulated for the entire functional range of items.

RESULTS

Scaling the Data Sets

Both the observed and simulated data were scored using Winsteps (Linacre, 2005). The distribution of item calibrations and person ability estimates for both data sets are illustrated in Figures 1 & 2. For ability estimates, the results across datasets were comparable, but not identical. Both datasets contained 89,116 examinees, but the average number of items administered [Observed data = 107.8, Simulated data = 97.8] was a little different (Table 1); however, there were no test records that contained fewer than 60 or more than 250 items. This indicates that the minimum and maximum limits imposed by the algorithm was working correctly. The ability estimates generated [Observed data, mean = 0.51 (0.80); Simulated data, mean = 0.60 (0.93)] were also similar, but not identical. The two datasets produced comparable person separation indices, although the index for the simulated data was slightly higher because the standard deviation of the simulated dataset was larger.

Table 1. Comparison of Ability Estimates for Observed and Simulated Data

Real Data

SUMMARY OF 89116 MEASURED Examinees VALID RESPONSES: 5.5%

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	55.2	107.8	.51	.23	1.00	-.1	1.01	-.1
S.D.	35.5	70.8	.80	.06	.06	1.0	.11	1.0
MAX.	147.0	250.0	4.08	.39	1.54	4.5	5.58	5.4
MIN.	11.0	60.0	-3.04	.13	.76	-4.0	.67	-4.0

REAL RMSE	.24	ADJ.SD	.76	SEPARATION	3.21	Examin RELIABILITY .91
MODEL RMSE	.23	ADJ.SD	.76	SEPARATION	3.26	Examin RELIABILITY .91
S.E. OF Examinee MEAN = .00						

Simulated Data

SUMMARY OF 89116 MEASURED Examinees VALID RESPONSES: 4.9%

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	50.2	97.8	.60	.23	1.00	-.2	1.01	-.1
S.D.	33.1	65.3	.93	.05	.05	.9	.11	.9
MAX.	151.0	250.0	4.53	.43	1.46	4.3	4.88	6.3
MIN.	8.0	60.0	-3.54	.13	.76	-4.0	.73	-3.8

REAL RMSE	.24	ADJ.SD	.90	SEPARATION	3.74	Examin RELIABILITY .93
MODEL RMSE	.24	ADJ.SD	.90	SEPARATION	3.78	Examin RELIABILITY .93
S.E. OF Examinee MEAN = .00						

For item difficulty calibrations, the results across datasets were also comparable, but not identical. The observed dataset contained only 1,973 items while the simulated dataset contained 2,000. This occurred because 27 items in the observed dataset were "turned off" before being administered (Table 2).
The difficulty calibrations generated [Observed data, mean = 0.00 (0.99); Simulated data mean = 0.00 (1.02)] were nearly identical[3]. The two datasets produced nearly identical item separation indices.

[3] Note that a mean of 0.00 for both item sets is not an empirical finding. Most Rasch calibration software defines the mean item calibration as 0.00 unless told otherwise. It is a common convention for data sets that are not being equated to some other frame of reference.

Table 2. Comparison of Item Calibrations for Observed and Simulated Data

Real Data

SUMMARY OF 1973 MEASURED Items LACKING RESPONSES: 27 Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	2493.8	4867.3	.00	.04	1.00	-.5	1.01	.1
S.D.	1731.5	3350.2	.99	.02	.03	3.0	.06	3.1
MAX.	8953.0	12213.0	3.99	.14	1.26	9.9	1.58	9.9
MIN.	83.0	370.0	-3.31	.02	.78	-9.9	.71	-9.9

REAL RMSE	.05	ADJ.SD	.99	SEPARATION	21.15	Item	RELIABILITY 1.00
MODEL RMSE	.05	ADJ.SD	.99	SEPARATION	21.31	Item	RELIABILITY 1.00

S.E. OF Item MEAN = .02

Sim Data

SUMMARY OF 2000 MEASURED Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	2235.6	4356.2	.00	.04	1.00	-.6	1.01	.0
S.D.	1359.3	2731.8	1.02	.02	.02	1.9	.05	2.1
MAX.	5351.0	10089.0	3.43	.16	1.11	7.0	1.36	8.2
MIN.	158.0	278.0	-3.24	.02	.93	-5.3	.82	-4.6

REAL RMSE	.05	ADJ.SD	1.01	SEPARATION	21.78	Item	RELIABILITY 1.00
MODEL RMSE	.05	ADJ.SD	1.01	SEPARATION	21.88	Item	RELIABILITY 1.00

S.E. OF Item MEAN = .02

PCA of NCLEX residuals

A principle components analysis was performed on the standardized residuals from both datasets. Although data sets of this size can be calibrated and analyzed with regard to fit, displacement, and the like, rather quickly, PCA takes much longer. The results for the two datasets are summarized in Table 3. Although the datasets were not identical, the differences do not seem large enough to degrade the quality of the conclusions drawn. Across both datasets, the largest factor (factor 1) accounted for less than one fifth of one percent of the total residual variance. With a first factor that is this small, it is nearly impossible to argue that there is any noticeable structure at all.

This is good news for NCLEX, but it does not make for a good example to illustrate the need to identify the difference between a real first factor and an artifact of the data. Here, the reader can dismiss the possibility of another dimension, without ever considering how much of the first factor is just an artifact of the data. No one would care about a factor of this size even if it did exist. It is akin to arguing that some people are paying more at the butcher shop for their meat because they request that the butcher use a slightly heavier grade of wrapping paper.

Despite the small size of the first factor, one may note that the first factor for the observed data is larger than the first factor of the simulated data. One could argue that the first factor of the simulated data is essentially a baseline and that only factors larger than that can have meaning. In this case, the difference would account for approximately one tenth of one percent of the total residual variance. This difference is also quite small and generally confirms the unidimensionality of the NCLEX-RN.

Table 3. Comparison Summary of Observed and Simulated Data.		
	Observed Data	**Simulated Data**
Candidates	**89,116**	**89,116**
Items[1]	**1,973**	**2,000**
Total Residual Variance[1] **(in Eigenvalue units)**	**1,973**	**2,000**
Factor 1	**3.5 (0.18%)**	**1.4 (0.07%)**
Factor 2	**2.1 (0.11%)**	**1.4 (0.07%)**
Factor 3	**1.9 (0.10%)**	**1.4 (0.07%)**
Factor 4	**1.8 (0.09%)**	
Factor 5	**1.8 (0.09%)**	

Note: The largest factor (Factor 1) accounted for less than one fifth of one percent of the total variance in the residuals.

[1] In the observed data, 27 items were turned off and therefore not administered to any candidates. Ideally, the number of items and candidates in the simulated data should match the observed conditions exactly, but this minor difference should not substantially harm the interpretability of the results.

DISCUSSION

The Rasch model requires that there is a single dimension, only the difference between B_n and D_i matters. However, in combining several content areas into the general construct of nursing ability, there are concerns that there could be some multidimensionality. Rather than modeling it in a multidimensional model, the choice was made to hold it constant. That is the basis for our test plan specifications. As a result, we have controlled the multidimensionality to prevent vast difference from person to person.

The advantages of this method of testing for multidimensionality include simplicity in communication and the ability to accommodate sparse data matrices that are not missing at random. Methods that are sufficient and easy to communicate are important. A comparison of observed structure with ideal permits the less technically inclined reader to understand the comparison without having to be conversant in factor analysis.

The disadvantages of this method are primarily the laborious nature of adequately simulating the data and the amount of time that it takes to run PCA on a data matrix of this size. However, there are also some limitations that are attributable to the nature of the data. In an adaptive test, there are very few off-target items. As a result, no response is terribly unexpected, which makes it difficult to identify misfit to the model or multidimensionality. Therefore the conclusion that degree of multidimensionality is practically zero, should be treated somewhat skeptically. Although this sparse dataset does not indicate any multidimensionality, it is possible that a complete data matrix would. Similar investigations with untargeted pretest data could provide a test for how data missing at random would fall out.

Future enhancements could include other confirmatory methods to demonstrate that the factors make no difference. Also, if the same type of matrix will be routinely examined, it may be practical to run several simulation datasets to assess baseline and then use that same baseline to examine future observed datasets. One could dispense with the simulation portion after the baseline is well established. Comparisons with other fit analyses may also prove to be enlightening.

REFERENCES

Guilford, G. P. (1936). *Psychometric Methods*. New York: McGraw-Hill.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment test. Chicago: MESA Press. (Original work published 1960).

Linacre, J. M. (2005). Winsteps version 3.50. Rasch Measure Calibration software.

Figure 1. Observed Data

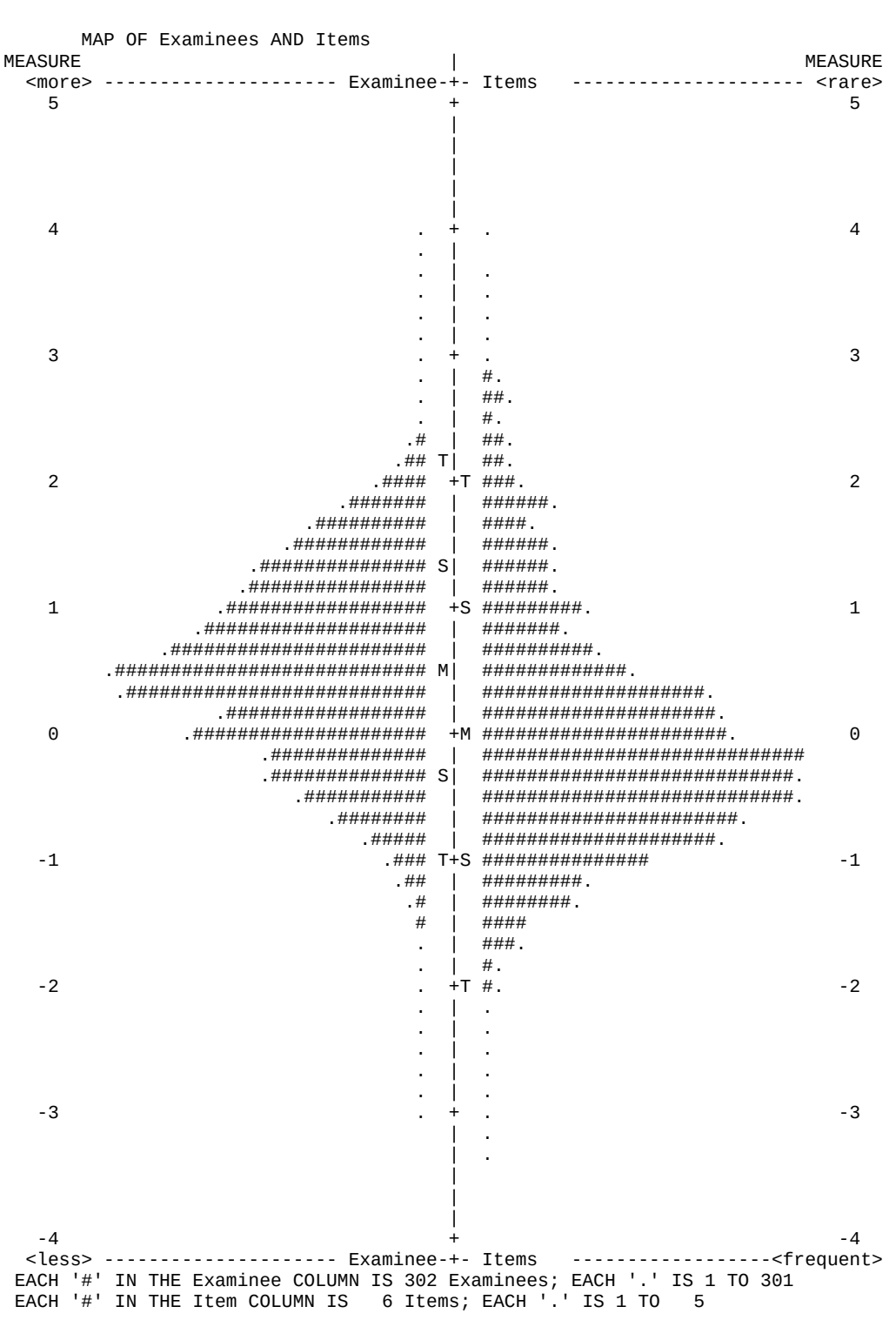

Figure 2. Simulated Data