DOCUMENT RESUME

ED 049 305 TM 000 500

| | |
|---|---|
| AUTHOR | Durovic, Jerry |
| TITLE | Application of the Rasch Model to Civil Service Testing. |
| INSTITUTION | New York State Dept. of Civil Service, Albany, N.Y. |
| PUB DATE | Nov 70 |
| NOTE | 11p.; From symposium "Application of the Rasch Model to Test Development" presented at the Annual Convention of the Northeastern Educational Research Association, Grossingers, New York, November 1970 |
| EDRS PRICE | EDRS Price MF-$0.65 HC-$3.29 |
| DESCRIPTORS | Computer Programs, Goodness of Fit, Government Employees, *Item Analysis, *Models, *Personnel Evaluation, Personnel Selection, Scores, *Test Construction, *Testing, Tests |
| IDENTIFIERS | Rasch Model |

ABSTRACT

The New York State Department of Civil Service investigated an empirical application of the Rasch model which appears useful in Civil Service testing. The model is a powerful tool for developing insights into what test items are measuring while permitting the investigator to spot defective items. It also reveals meaningful distinctions in the type of task set by different items. Two sets of specific examples are discussed to illustrate its usefulness. The first set of examples considers the use of item probability as an index of the degree of fit of the material to the model, while the second discusses the "normal deviate" matrix, which displays the goodness of fit of each item at each score group, and enables an investigator to ascertain the overall validity of the general index. These examples demonstrate the applicability of the Rasch model to a variety of conditions. The author suggests that the model seems promising for civil service testing since it is not simply a means to derive scores but is also a powerful tool for test analysis, construction, and design. (CK)

# APPLICATION OF THE RASCH MODEL TO CIVIL SERVICE TESTING*



Jerry Durovic
Associate Personnel Examiner
New York State Department
of Civil Service
State Office Building Campus
Albany, New York

(518) 457-2374



*Presentation for November 16, 1970 NCME Symposium on Application of the Rasch Model to Test Development

ANNUAL CONVENTION OF THE

NORTHEASTERN EDUCATIONAL RESEARCH ASSOCIATION

GROSSINGERS, NEW YORK
NOVEMBER, 1970

# APPLICATION OF THE RASCH MODEL TO CIVIL SERVICE TESTING

A Paper Prepared for the Symposium on
Application of the Rasch Model to Test Development

By
Jerry Durovic
New York State Department of Civil Service

## Introduction

At the New York State Department of Civil Service we have been working exclusively with only one of the models[1] developed by Rasch. Our interest has been focused on the model discussed by Wright at the 1967 ETS Invitational Testing Conference.[2] This particular model is occasionally referred to as the log-odds model and basically treats the responses to a test item in terms of dichotomies such as right and wrong.

Our testing program places a heavy emphasis upon the multiple-choice type of written test item. In general we use 4 or 5 choice items which are scored as right or wrong. These test items are generally grouped into sets of 15 or 20 questions, which we call subtests, which are designed to measure some particular area. We have analyzed a variety of our subtests with the Rasch Model since we began working with it. We have analyzed subtests in:

| | |
|---|---|
| Abstract reasoning | Quantitative reasoning |
| Reading comprehension | Statistics |
| Supervision | Administrative judgment |
| Economics | Budgeting |
| Vocabulary | Interviewing techniques |
| Spelling | Report writing |
| Spatial relations | |

to name several. Some of our subtests might be called aptitude tests and others achievement tests by educators. Unfortunately the public personnel selection community and the educational community occasionally use different labels for similar ideas. As a result I will try to limit my use of labels in order to avoid misunderstanding.

---

[1] Rasch, G. Probabilistic models for some intelligence and attainment tests, Copenhagen: Danish Institute for Educational Research, 1960. Chapters V-VII, p.

Objective

In our empirical work with the Rasch Model we have been exploring a variety of issues simply because we understand very few of them and have virtually no recourse to published works since they are relatively scarce.

One of the issues we have been exploring is one that I would like to relate to you today. It is an empirical application of the model that has aroused our enthusiasm, appears to be particularly useful for civil service testing and nevertheless is one that we have not read much about in the literature.

Our empirical investigation of the Rasch Model on test items, measuring a wide variety of areas, has led us to believe that the Rasch Model may be a powerful tool for developing insights into what test items are measuring. We believe further that this information can then be used to make practical decisions about item construction and/or test design.

Specifically our experience has shown us that the Rasch Model permits us to

1. spot defective items, in the traditional sense such as items without good key answers; and
2. spot items which do not belong in a subtest in the sense for example, of items which are functioning as reading comprehension items or quantitative reasoning items in subject matter tests.

We have found repeated instances of the model quickly and clearly pointing up meaningful distinctions in the type of task set by different items.

It would be unreasonable for me to expect to convince you, within the limited span of time I have today, of this general applicability of the Rasch Model which we have repeatedly found and believe [redacted] may be an important aspect of the Rasch Model.

2. Wright, B. "Sample-free test calibration and person measurement." In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1968, 111, 85-101.

What I propose to do in the time remaining however, is to discuss various specific examples of our work which have sparked our enthusiasm, and helped formulate our current view in the belief, that for some of you, this may be a useful application, and in the hope that you may become sufficiently interested to explore these aspects more carefully and systematically with your own material, in order to nail down this particular application, if in fact it is a real one.

## Item Probability

Example #1:

One of our earliest empirical investigations came about as a result of one of those misunderstandings that occasionally occur between the public personnel selection sector and the educational community. At the first AERA presession on the Rasch Model conducted by Dr. Wright a few years ago in California, the participants were encouraged to bring data with them to be analyzed at the pre-session.

One of our colleagues from the United States Civil Service brought along data on one of their widely used tests. His data was in alphabetic form, that is the alternatives to test items were A, B, C, etc., rather than numeric or 1, 2, 3, etc. Unfortunately, the programs used for the analysis at UCLA for the presession were designed to accept only numeric data. (Parenthetically I might note that the data I brought to the same session suffered the same fate. In fact the data brought by those in the educational community were only numeric while those from public personnel agencies brought only alphabetic data and neither thought to question this point until it was too late.) Fortunately, we had a working computer program at the New York State Department of Civil Service which could handle both alphabetic and numeric data and we agreed to analyze the U. S. Civil Service data when we returned.

Their data was selected by them for analysis because it had been reviewed over the years by traditional analysis and it was felt that this set of items could reasonably be expected to fit the assumptions of the Rasch Model and therefore could serve as a useful means of gaining a better understanding of the Model.

The data consisted of the responses of 987 middle managers from various agencies in the Federal Government to a set of 25 Reading Comprehension items. We submitted this data to an analysis by the Rasch Model and examined one of the indices which appears as a routine output on our program and which is used to determine the degree of fit of the material to the model. The index used is called the "item probability" and is the probability of the observations, given the item fits the model. For each test item an "item probability" is calculated. Glancing at the values of the "item probabilities" for this data we noticed several items whose "item probabilities" were extremely low in comparison to the other items. In particular we noticed four items with "item probabilities" of 0.000.

So, we looked at the content of all the items and found that the items could be answered correctly based solely on the information presented in the stem, except for the four items with the low "item probabilities". These four items required the candidate to bring additional knowledge to the item in order to correctly select the key answer, a requirement that was not present in the other items.

Next, we looked at the "item probabilities" of all 25 items again and noticed three items with relatively high "item probabilities". A look at the content of these items revealed a dramatic shift in the vocabulary level in these items, as compared to the other 22 items. These three items contained phrases such as "organized acquisitive procedure for self-maintenance" as a definition of war; and "non-pecuniary incentive."

The use of the "item probabilities" in this way, for this set of data, seemed to enable us to spot items that were functioning differently than the total set of

items and we seemed to be able to find reasonable explanations of the nature of the functioning differences.

Example #2:

In another instance, we examined a set of test items that we developed as part of a promotion examination for second level professional research positions. We call this examination the Research Services series. For one of these positions we developed a set of 30 questions in the area of social research. The items were all designed to measure the same general area. For purposes of analysis the 30 questions were treated as two blocks of 15 items. Each block of 15 items was subjected to a Rasch analysis (N=261) and both fit the model ($p = 0.474$, and $p = 0.416$, respectively).

For the first set of 15 items we again looked at the "item probabilities." Four items had "item probabilities" that were relatively high in comparison to the rest of the items. These four items had probabilities greater than 0.64 while the remaining eleven items were all below 0.45. In keeping with our prior experience we hypothesized that the four items with the relatively high "item probabilities" had some kind of common feature that could explain their similar extreme probabilities. After examining the content of each of the 15 items it appeared as if we could identify a concept that set these four items apart. These four items were unique from the rest of the subtest in that these items seemed to require a person to be aware of the practical considerations as well as the theoretical and ideal approaches to research. The other 11 items did not have this requirement.

To explore this hypothesis further we looked at the second set of 15 items and found two items which seemed to embody the same concept that we found in the four items of the first set. A check of their "item probabilities" showed that these two items had the lowest probabilities of fit on the entire subtest, that is, these two items were deviant from the rest of the set of items in their "item bilities" and in a similar manner. Thus the Rasch analysis seemed to indicate

items with extreme "item probabilities" embodying a common concept.

## Normal Deviate

There is an alternative method of evaluating the fit of items which involves an examination of the "normal deviate" matrix. This matrix displays the goodness-of-fit of each item at each score group. It enables us to look for patterns as well as locate the precise point of poor fit if it exists. This often helps us to determine if a general index of poor fit is important to consider or not. For example, if the overall fit of the item is poor as reflected in a general index, but an examination of the "normal deviate" matrix shows us that the fit is poor only at one score group and if that score group is an unusually small one, we are likely to overlook the general index or at least not pay too much attention to it. The "normal deviate" matrix possesses a general index of fit too called the "mean square fit". The "mean square fit" and the "item probability" are related to each other and either can be used. Some of the examiners in our agency prefer one, while others prefer the other. While either can be used, the personal preferences of the examiners performing the analysis are factors to be considered, especially if one has to work with them on a daily basis. In order to preserve a degree of harmony among my colleagues I intend to now give "equal time" to the "mean square fit" advocates.

Example #1:

You may recall that the last examination I was discussing was the Research Services examination, which was a promotion examination for second level professional research positions. Another part of that examination was a subtest on "tabular interpretation". For this subtest, the candidate is required to recognize relationships in and draw inferences from data presented in tabular form. The items stress the ability to deal with relationships among the various categories in the tables but minimizes the need to perform computations to arrive at the

For this material the analysis with the Rasch Model was a bit more involved. We have three alternate forms of the Research Services examination; Forms A, B, & C. Each form contains a subtest on "tabular interpretation". We use a different 15 item "tabular interpretation" subtest on each Form, however each subtest contains some items common to at least one of the other subtests.

We analyzed the subtest from Form A (n = 275) with the Rasch Model. One item possessed an extreme "mean square fit" of 3.25 as compared to the rest of the items which had "mean square fits" of less than 2.00.

We then examined the item and seemed to find:

1. the item required a successful manipulation of a series of complex relationships which was not required by the other items, and
2. one of the wrong answers could be arrived at by performing all but the final operation.

We felt that perhaps many of the better candidates understood the item but were omitting the final operation and therefore getting the item wrong.

Fortunately, we were in a position to evaluate the tentative conclusions. Form B, an alternate Form for this examination also contained this item, which for purposes of our discussion I will call item X. In Form B, a different item preceded item X, and this new item could be answered correctly by performing all but the final operation required for item X. It was expected that the better candidates would get this new item correct and then would realize that the answer to item X simply required an additional operation. Rasch analysis of Form B (n = 188) supported our expectation. Both the new item and item X had "mean square fits" slightly below 1.00.

In examining the "normal deviate" matrix for Form B we also found a single item with an extremely high "mean square fit" (MSF = 7.89). A review of all the items revealed that this particular item was unique in that the candidate was required to deal with an indeterminacy to arrive at the correct answer. In the other items the answer could be found by simply filling in all missing entries in the table. It was felt, that manipulating all possible combinations through an indeterminacy may require a higher level of reasoning ability then simply filling in missing entries. This item was also included in Form C (n = 248) and again was the only item with a high "mean square fit" (MSF = 4.93). The Rasch analysis seemed to again pinpoint the item that was functioning in a manner different from the rest of the subtest.

Example #2:

I would like to discuss one final example. We have not limited our investigations to tests for positions that require a college education as the previous examples might lead you to believe. For example, let us look at an examination for Clerks. As part of our promotion examination to second level clerical positions we include a subtest of 15 items on vocabulary. The items in this subtest present five alternatives from which the candidate is to select the one most similar in meaning to the word given in the stem of the item.

We subjected the 15 vocabulary items to a Rasch analysis (n = 3,624) and found one item with a high "mean square fit" of 17.96. An inspection of this item revealed:

1. for the word presented in the stem, it was possible to derive an entirely different word by changing only one or two letters, and
2. the meaning of this new word was one of the incorrect alternatives that could be chosen by the candidate.

In another vocabulary subtest for an alternate form of this examination we found two items with similar properties. A Rasch analysis of the data (n = 3,998) showed that these two items also had high "mean square fits".

## SUMMARY

There are many other examples that I could present to you but I believe the ones that have been presented here today give you an idea of the application that we see as one of the most promising for civil service testing. I discussed items in reading comprehension, social research, tabular interpretation, and vocabulary. I have attempted to present items from a variety of areas to give you an idea of the general applicability we have found.

I discussed examples of single subtests, multiple subtests with common overlapping items, as well as separate subtests with identical item construction concepts. I have tried, in this way, to illustrate the persistence of our finding across a variety of conditions.

As for our subjects, they were both college as well as non-college personnel. They included professional technical specialists, professional managers or generalists, and clerical or non-professional personnel. In this way I have attempted to give you an idea of the wide subject levels that we find succumb to the Rasch Analysis.

Within this brief period that I have had with you I have attempted to illustrate that, despite wide variations of item content and subject composition, the Rasch analysis produced persistent useful results.

Our feeling about the applicability of the Rasch Model to civil service testing perhaps is best summarized by a comment made by Dr. Albert P. Maslow, Chief of the Personnel Measurement Research and Development Center, of the United States Civil Service in reviewing an article[3] on the Rasch Model to be published in an upcoming issue of the Public Personnel Review:

[3] Durovic, J. "Improving Written Test Measurement in Personnel Selection." Public Personnel Review, in print.

"I believe that a major value of the Rasch Model may prove not to be simply as a means to derive scores, but as a powerful tool for developing insights about what it is that test items are measuring; how the items in a test relate to one another, and how these insights can apply to the improvement of test design, item construction, and test analysis."