1993

Measurement precision test construction and best test design

Pender J. Pedler

FREEDOM THROUGH KNOWLEDGE

EDITH COWAN UNIVERSITY

PERTH WESTERN AUSTRALIA

MEASUREMENT PRECISION TEST CONSTRUCTION AND BEST TEST DESIGN

Pender J. Pedler

RESEARCH REPORT No: 7

December 1993

MEASUREMENT ASSESSMENT
and
EVALUATION LABORATORY

ACKNOWLEDGEMENTS

The author is grateful for the assistance offered by
Professor David Andrich, Murdoch University and
Dr Barry Sheridan, Edith Cowan University
during the conduct of the research featured,
and for a critical review of the draft revision of,
this report

Measurement Precision, Test Construction and Best Test Design

Abstract

This article examines the precision of measurements obtained from using the Rasch Dichotomous Model to analyse test data. Considering tests in which the item difficulties are uniformly spaced from easiest to most difficult, permits the derivation of an alternative expression for the standard error of measurement. This expression is sufficiently simple to enable the precision properties of uniform tests to be readily described and to enable a variety of problems of test construction to be solved. One particular problem is that of best test design. Regarding measurement precision as a property of the test only, we show that the best uniform test of a given length and a given target interval is the one that satisfies a minimax condition on the standard error. We illustrate the solution to this problem and describe properties of best tests.

Key words: Rasch model, standard error, test construction, best test design.

Measurement Precision Test Construction and Best Test Design

1. Introduction

The precision of a measurement is its degree of accuracy or exactness. All measurements are imprecise, each subject to an error due to the limitations of the measuring instrument used. A social scientist administering an educational or psychological test is measuring a person or set of persons with respect to the variable defined operationally by the test items. Thus the measurements obtained by the scientist are likewise imprecise, each subject to an error due to the limitations of the test itself and the limitations in transforming the test data into measurements.

Item response theory assumes that the observed data can be regarded as the outcome of an item response model, a statistical model of all possible responses of all persons to the test items. A particular set of test data is then transformed into a measurement on each person by obtaining the best estimate of the model parameter for each person consistent with the test data. Each measurement is thus a statistical estimate and subject to a random or statistical error. The extent of this error is usually given by the value of the standard error. Questions concerning the precision of the measurements obtained from using the test can then be answered by an appropriate calculation and interpretation of the standard error of measurement.

The simplest response format records only two levels of performance on each test item. These levels are usually referred to as 'wrong', 'incorrect' or 'fail' scored level zero (0), and 'right', 'correct' or 'pass' scored level one (1). Test items with this response format are termed dichotomous items. Most of the literature of item response theory is concerned with models for the analysis of dichotomous test data. Of this literature on dichotomous models, a significant proportion is concerned with the Rasch Dichotomous Model (RDM). Rasch (1960) introduced this model to analyse a set of data arising from a military group intelligence test. Since then, it has been widely applied to the analysis of test data and to the development of item banks (see, for example, Wright & Stone, 1979).

2. The Rasch Dichotomous Model

The Rasch Dichotomous Model (RDM) specifies that the dichotomous response x_{ni} of the person n; $n = 1, 2, ..., N$; to the test item i ; $i = 1, 2, ..., L$; depends on the value of the ability parameter β_n of the person n and the difficulty parameter δ_i of the item i only through the value of their difference $\theta_{ni} = \beta_n - \delta_i$. These parameter values are real numbers and thus may be represented as points on a continuous line or continuum as in Figure 1. This line represents the test variable and shows the relationships between persons and items with respect to the test variable.

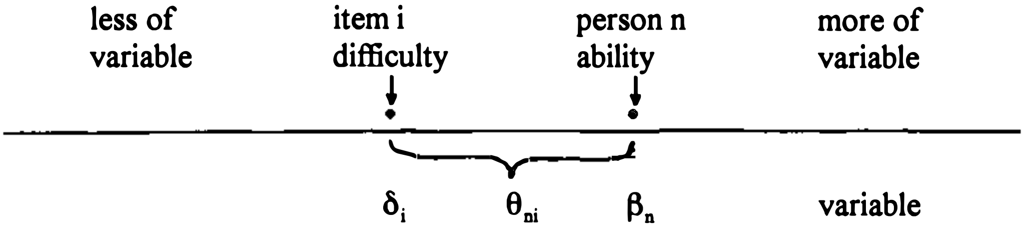


Figure 1: The test variable together with persons and items.

The dichotomous response x_{ni} of the person n to the item i is modelled with the probability function.

$$\begin{aligned} P(x_{ni}\ ;\ \beta_n, \delta_i) &= e^{x_{ni}\theta_{ni}} / (1 + e^{\theta_{ni}})\ ; \\ \theta_{ni} &= \beta_n - \delta_i\ ; \\ x_{ni} &= 0,\ 1. \end{aligned} \tag{1}$$

As the difference $\theta_{ni} = \beta_n - \delta_i$ is in the exponent of an exponential term, the scale unit of the variable is a logarithm unit, usually contracted to logit.

Dropping the subscripts n and i designating the particular person and the particular item, the expected value of the response of the person to the item, as a function Φ of the person-item difference $\theta = \beta - \delta$, is given by

$$\begin{aligned} \Phi(\theta) &= E[\,x\ ;\ \theta = \beta - \delta\,] \\ &= P(\,x = 1\ ;\ \theta = \beta - \delta\,) \\ &= e^{\theta} / (1 + e^{\theta}) \end{aligned} \tag{2}$$

The function Φ is referred to as the characteristic function of the RDM while the graph of Φ against θ is referred to as the characteristic curve.

The information I about the value of the person parameter β contained in a single response x of the person to an item is then (Lord, 1980, 70-73; Rao, 1973, p.309; Wright & Stone, 1979, p.135) given by

$$I(\theta) \;=\; E[\ \{ d\ln P / d\theta \}^2 \ ;\ \theta = \beta - \delta\ ] \;=\; \Phi'(\theta) \;=\; e^{\theta}/(1+e^{\theta})^2 \quad (3)$$

The information I specifies the extent to which uncertainty concerning the unknown value of θ and hence that of the person parameter β, is reduced as a consequence of an observed response of the person to a single test item. The greater the value of this information I, the greater is the reduction in uncertainty and the more precise is the subsequent measurement for β.

The value I (θ) of the item information function I is given by the rate of change $\Phi'(\theta)$ of the characteristic function Φ for each value of the person item difference $\theta = \beta - \delta$. The graph of I against θ, the item information curve of the RDM, is unimodal with horizontal asymptote I = 0. The graph is symmetric about $\theta = 0$ with a maximum turning point at $\theta = 0, I = \frac{1}{4}$. Thus maximum information I about the value of the person parameter β arising from an observed response of the person to a single item, occurs when the item difficulty δ is equal to the person ability β. We interpret such an item as being targeted on the person.

The information J about the value of the person parameter β contained in a set of L responses is obtained by summation. Under the assumption of local independence, the responses of the person to each of the test items are independent variables and the information in each response contributes additively to the total test information. Thus

$$J \;=\; J(\beta; \delta_1\delta_2, \ldots, \delta_L) \;=\; \sum_{i=1}^{L} I(\theta_{ni})\ ; \quad (4)$$

$$\text{where} \quad I(\theta) \;=\; e^{\theta}/(1+e^{\theta})^2$$
$$\theta \;=\; \theta_{ni} = \beta_n - \delta_i = \beta - \delta_i.$$

The test information J about the value of the person parameter β specifies the precision of the subsequent measurement $\hat{\beta}$ of β as a consequence of a set of observed responses of the person to the test. The greater the value of J the greater is the precision of measurement and the smaller is the likely error between the measurement $\hat{\beta}$ and the unknown value of the parameter β.

The error of measurement is the difference $\hat{\beta} - \beta$ between the measurement $\hat{\beta}$ of the person parameter and the unknown value β. In practice, we have a single measurement $\hat{\beta}$ for each person arising from the analysis of the particular set of test data. As the true value β of the person parameter is unknown, that of the error $\hat{\beta} - \beta$ is likewise unknown. However, we may conceive the RDM as modelling infinitely many responses of the person to the test, giving not just a single measurement $\hat{\beta}$ but a distribution of infinitely many possible values. This distribution may be referred to as the sampling distribution of the measurement $\hat{\beta}$, and depends on the model and the estimation procedure used to obtain the measurement $\hat{\beta}$ from the test data.

Maximum likelihood estimation theory (Anderson, 1980, p.60; Habermann, 1977; Swaminathan, 1983, p.30) ensures that the sampling distribution of the measurement $\hat{\beta}$ has asymptotic normal distribution with mean β, variance J^{-1} and standard deviation $J^{-1/2}$. It follows that the random error $\hat{\beta} - \beta$ is asymptotically normal with mean zero and standard deviation $J^{-1/2}$. Thus the likely size of the random error is specified by the standard deviation $J^{-1/2}$, referred to as the standard error (SE) of the measurement $\hat{\beta}$. It follows from (4) that

$$\mathrm{SE} \;=\; \mathrm{SE}\,(\beta;\delta_1\delta_2,\ldots,\delta_L) \;=\; \left\{ \sum_{i=1}^{L} I\,(\theta_{ni}) \right\}^{-1/2}, \qquad (5)$$

where $\theta_{ni} = \beta_n - \delta_i \equiv \beta - \delta_i$.

The smaller the value of the standard error of measurement SE for β, the greater is the precision of the measurement $\hat{\beta}$ of β and the shorter are the confidence intervals $\hat{\beta} \mp z\,\mathrm{SE}$ for β using critical standard normal z values.

In practice, the standard error of measurement $\mathrm{SE}\,(\beta;\delta_1\delta_2,\ldots,\delta_L)$ is estimated by substituting the estimates $\hat{\beta}, \hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_L$ for $\beta, \delta_1, \delta_2, \ldots, \delta_L$ respectively in (5) and evaluating $\mathrm{SE}\,(\hat{\beta};\hat{\delta}_1,\hat{\delta}_2,\ldots,\hat{\delta}_L)$. Thus the standard error SE is a property of the measurement $\hat{\beta}$ and the test specified by the item difficulties $\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_L$. Its value does not depend on the parameter values β of other persons taking the test, nor on the fit or lack of fit between the data and the item response model. However, the focus of this study is on the standard error of measurement as given by (5) rather than its estimated value from a particular data set.

Theoretically, the value of the person parameter β is merely an arbitrary point on the continuum representing the test variable. Thus the standard error SE may be regarded as a property only of the test, indicating the degree of precision of a potential measurement $\hat{\beta}$ at any point β, rather than just those obtained from a particular data set. Rasch (1960) appreciated this point well when he stated that

> the precision or reliability of a test is conceived as an intrinsic property of the test - not as a property of the "population", or any other collection of testees, to which the test has been applied on some occasion (p.33).

Thus the target of the test is merely a set of possible values of the parameter β. No assumption concerning the distribution of these values need be made. Like the process of item calibration, measurement precision may therefore be described as being 'distribution-free'. Although the target of the test may potentially be the whole continuum $\{ \beta \mid -\infty < \beta < \infty \}$, it is more usual to consider a finite target interval

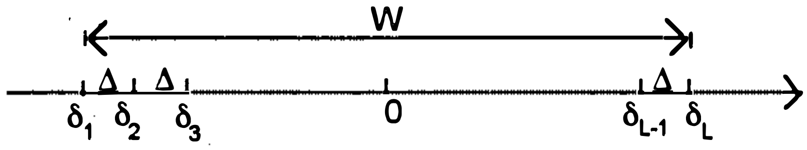
$$B = [-\beta_0, \beta_0] = \{\beta \mid -\beta_0 \le \beta \le \beta_0\}$$

for some finite value β_0 of β.

Finally, we note that (5) shows that the precision of measurement achieved by the test depends on the test only through the values $\delta_1\ \delta_2, \ldots, \delta_L$ of the set of item difficulties. However, before we use this to describe precision properties of tests and to solve problems of test construction, we make a simplifying assumption about the distribution of these item difficulties.

3. Uniform Dichotomous Tests

We now consider tests comprising a set of dichotomous items in which the item difficulties are uniformly spaced from easiest to most difficult. We choose this simplifying assumption of the distribution of item difficulties for two reasons. First, the spread of item difficulties in practice often can be approximated by a uniform distribution. Secondly, the experimental evidence on test construction suggests that, although information maximisation depends on both the distribution of persons and that of items, the uniform test is either the best or equivalent to the best for all practical purposes. Wright and Stone (1979, p.134), in summarising the evidence, recommend that "the best all purpose test is the uniform test".



Consider the L item difficulties $\delta_1\ \delta_2, \ldots, \delta_L$ ordered from easiest δ_1 to most difficult δ_L. With uniformly spaced item difficulties, the test is then specified by the test length L given by the number of dichotomous items, the test width $W = \delta_L - \delta_1$ being the range of the L item difficulties, and the item spacing $\Delta = W/(L-1)$ between consecutive item difficulties. These three test attributes are related as each can be expressed in terms of the other two as follows.

$$\Delta = W/(L-1) \quad ; \quad W = \Delta(L-1) \quad ; \quad L = 1 + W/\Delta . \tag{6}$$

As the sum of item difficulties is usually set to zero, we set the centre of the uniform test $(\delta_1 + \delta_L)/2$ as zero. It follows that

$$-\delta_1 \quad = \quad \delta_L \quad = \quad W/2 \quad = \quad \Delta(L-1)/2$$

and hence, for i = 1, 2, ..., L, that

$$\delta_i \quad = \quad \delta_1 + \Delta(i-1) \quad = \quad \Delta(2i-L-1)/2$$

Substituting now for $\delta_1\ \delta_2, \ldots, \delta_L$ into (4) and (5), it follows that the precision of the uniform test at a point β is now a function of the value of β and any two of the three test attributes, length L, width W and spacing Δ. However the constant item spacing Δ permits us to exploit the connection between the item information function I and the characteristic function Φ.

As the item spacing Δ is constant, the product of the test information J with the item spacing Δ, namely

$$J\,\Delta \quad = \quad \sum_{i=1}^{L} (\theta_{ni})\,\Delta$$

where

$$\theta_{ni} \quad = \quad \beta - \delta_i \quad = \quad \beta - \Delta(2i - L - 1)/2 \ ,$$

may be represented graphically. This product is the sum of the areas of L contiguous rectangles on the item information curve of the RDM as in Figure 2. The ith rectangle has base Δ centred at θ_{ni} and height $I(\theta_{ni})$.

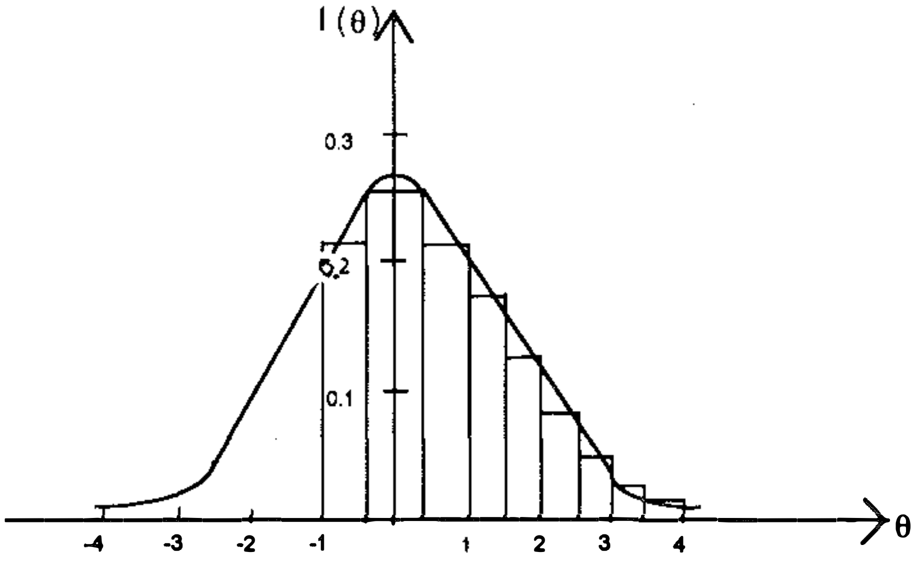


Figure 2: The product of the information J with the spacing Δ.

The particular example illustrated in Figure 2 has $L = 11$, $W = 5$, $\Delta = 0.5$, $\beta = 1.75$ and $\theta_{n1}, \theta_{n2}, \ldots, \theta_{nL} = 4.25,\ 3.75,\ \ldots,\ -0.75$, respectively.

Provided the item spacing Δ is small, the sum of the areas of these rectangles may be approximated by the area below the item information curve between appropriate limits. As the item information function I is the derivative Φ′ of the characteristic function Φ, this area has the following simple form.

$$\int I(\theta)\, d\theta \quad = \quad \int \Phi'(\theta)\, d\theta \quad = \quad \Phi^{+} - \Phi^{-}$$

where

$$\begin{aligned} \Phi^{+} &= \Phi^{+}(\theta_{n1} + \Delta/2) \\ \Phi^{-} &= \Phi^{-}(\theta_{nL} - \Delta/2) \end{aligned}$$

It follows that the test information J may be approximated as

$$J \quad = \quad \Delta^{-1} \int I(\theta)\, d\theta \quad = \quad \Delta^{-1}\ [\Phi^{+} - \Phi^{-}]$$

namely a scaled difference between two values of the characteristic function Φ. This expression is equivalent to that given by Douglas (1974, p.135, formula 32). We now derive an explicit expression for $\Phi^{+} - \Phi^{-}$ and hence J and SE, for the RDM.

Using

$$\theta_{n1} + \Delta/2 \;=\; \beta + \Delta(L-1)/2 + \Delta/2 \;=\; \beta + \Delta L/2 \;=\; \beta + \omega,$$

and

$$\theta_{nL} - \Delta/2 \;=\; \beta - \Delta(L-1)/2 - \Delta/2 \;=\; \beta - \Delta L/2 \;=\; \beta - \omega,$$

where

$$\omega \;=\; \Delta L/2 \;=\; (W + \Delta)/2$$

it is convenient to introduce ω as another test attribute. As ω is slightly greater than W/2, we term it the adjusted half width (AHW) of the test. Then

$$\begin{aligned}\Phi^{+} - \Phi^{-} &= \Phi(\beta + \omega) - \Phi(\beta - \omega)\\ &= e^{\beta+\omega}/(1+e^{\beta+\omega}) - e^{\beta-\omega}/(1+e^{\beta-\omega})\\ &= (e^{\omega} - e^{-\omega})/(e^{\beta} + e^{-\beta} + e^{\omega} + e^{-\omega})\\ &= \sinh\omega/(\cosh\beta + \cosh\omega)\end{aligned}$$

where

$$\begin{aligned}\sinh z &= (e^{z} - e^{-z})/2\\ \cosh z &= (e^{z} + e^{-z})/2\\ \tanh z &= \sinh z/\cosh z\end{aligned}$$

are the hyperbolic trigonometric functions.

It follows that the test information J and the standard error of measurement SE of a uniform test are given by

$$J \;=\; J(\beta) \;=\; J(\beta; L, W, \Delta, \omega) \;=\; \Delta^{-1}\sinh\omega/(\cosh\beta + \cosh\omega) \quad (7)$$

and

$$SE \;=\; SE(\beta) \;=\; SE(\beta; L, W, \Delta, \omega) \;=\; \{\Delta(\cosh\beta + \cosh\omega)/\sinh\omega\}^{1/2}$$

We thus have an explicit expression for the standard error of measurement SE of a uniform test in terms of standard scientific functions.

To illustrate, consider the calculation of the standard error SE for the particular example illustrated in Figure 2. Using (7), as $\beta = 1.75$, $L = 11$, $W = 5$, it follows that $\Delta = 0.5$, $\omega = 2.75$ and

$$\begin{aligned} SE &= SE(1.75) \\ &= SE(1.75;\ 11,\ 5,\ 0.5,\ 2.75) \\ &= \{\ 0.5\ (\cosh 1.75\ +\ \cosh 2.75) / \sinh 2.75\ \}^{1/2} \\ &= 0.8333 \end{aligned}$$

Using the earlier expression (5), the value of the standard error is found to be $SE = 0.8327$. Thus the discrepancy in using (7) rather than (5) for SE is 0.0006 or less than 0.1%.

The alternative expression (7) for the standard error of measurement of a uniform test gives virtually the same result as (5) for all practical purposes. We can see from Figure 2, that for a uniform test of a reasonable length L, the item spacing Δ is small and the two areas are virtually equal. A computer program was written to explore the discrepancy in standard error between the two expressions for uniform tests of length $L \geq 10$ and width $W \leq 10$ over the range $-4 \leq \beta \leq 4$ of the continuum. These bounds were chosen to be somewhat greater than would normally occur in practice. The maximum absolute discrepancy in standard error between (5) and (7) for uniform tests of different lengths is given in Table 1.

Table 1

Maximum absolute discrepancy in SE; $-4 \leq \beta \leq 4$, $W \leq 10$.

Test length L	Maximum discrepancy
$L \geq 10$	4.6×10^{-3}
$L \geq 20$	8.0×10^{-4}
$L \geq 50$	8.0×10^{-5}
$L \geq 100$	1.5×10^{-5}

These results demonstrate that, for all practical purposes, (7) gives the same result as (5) for uniform tests. However, (7) has the advantage of simplicity and provides a ready means of studying the precision of measurement with the RDM. In particular, we can examine how the test attributes such as test length and width affect the size of the standard error of measurement.

4. Measurement Precision with Uniform Tests

Questions concerning the precision of a test comprising uniformly spread dichotomous items can now be answered readily using (7). We illustrate this first with a numerical calculation, and secondly by examining how test attributes affect the precision of the test.

Example: *We have constructed a test with 45 items uniformly spread over a range of 4 logits. How precise are the measurements for persons*

(i) on target at the centre of the test?

(ii) 1 logit from the centre of the test?

(iii) 3 logits from the centre of the test?

Solution: As $L = 45$, $W = 4$, $\Delta = W / (L - 1) = 0.091$, $\omega = \Delta L/2 = 2.045$ and hence,

(i) $SE(0) = 0.343$

(ii) $SE(1) = 0.362$

(iii) $SE(3) = 0.579$

These calculations are readily performed on a scientific calculator, preferably a programmable one with the formula for SE stored as a program.

Consider now the measurement precision at the centre of a uniform test. This has been studied by Woodcock (1992) who used (5) to construct a Test Design Nomograph, graphs of SE against width W for tests of various lengths L and item spacings Δ. Putting $\beta = 0$ in (7), and using the identity that

$$\sinh \omega / (1 + \cosh \omega) = \tanh (\omega/2)$$

it follows that the test information J and the standard error of measurement SE at the centre of a uniform test are given by

$$J \;=\; J(0) \;=\; J(0;\, L,\, W,\, \Delta,\, \omega) \;=\; \Delta^{-1}\tanh(\omega/2) \qquad (8)$$

and

$$SE \;=\; SE(0) \;=\; SE(0;\, L,\, W,\, \Delta,\, \omega) \;=\; \{\, \Delta / \tanh(\omega/2)\, \}^{1/2}$$

This formula can now be used to replace all applications of the Woodcock Test Design Nomograph.

Although the precision at the centre of a test is determined by the values of both the item spacing Δ and the adjusted half width (AHW) ω, it is relatively insensitive to variation in ω provided ω is reasonably large. Specifically, as

$$0 \leq \tanh z \leq 1 \quad \text{for all } z > 0$$

and

$$\tanh z \rightarrow \text{ as } z \rightarrow \infty$$

it follows that

$$\begin{aligned} SE(0) &\geq \Delta^{1/2} & & (9)\\ &\approx \Delta^{1/2} \qquad \text{if } W \text{ is large ;} & & \\ \Delta &= W/(L-1)\,. & & \end{aligned}$$

We conclude that, when a test comprises a number of dichotomous items with item difficulties uniformly spread over a wide range, the precision at the centre of the test is essentially a function of the item spacing Δ only.

We now examine how the precision of a uniform test varies over the continuum. From (7), it follows that, for a given test, the test information function J is an even function of β. The graph of J against β, the test information curve, is symmetric about $\beta = 0$ with a maximum turning point at $\beta = 0$, $J = J(0)$. Furthermore, like the item information curve of the RDM, the test information curve is unimodal with horizontal asymptote $J = 0$. Similarly, the standard error curve, the graph of SE against β, is U-shaped, symmetric about $\beta = 0$ with a minimum turning point at $\beta = 0$, $SE = SE(0)$.

Consider now the precision of a test at each point β on the continuum relative to the precision at the centre of the test. Combining (7) and (9), it follows that the relative information and relative standard error at β is given by

$$J(\beta) / J(0) \quad = \quad (1 + \cosh \omega) / (\cosh \beta + \cosh \omega) \qquad (10)$$

and

$$SE(\beta) / SE(0) \quad = \quad \{ (\cosh \beta + \cosh \omega) / (1 + \cosh \omega) \}^{1/2}$$

From (10), we see that the precision of a test relative to the precision at the centre of the test depends on the test attributes only through the value of the AHW ω. Recall that as $\omega = (W + \Delta) / 2$, the relative precision is essentially a function of test width W only. The greater the width W of a uniform test, the greater is the relative information $J(\beta) / J(0)$ and the flatter is the test information curve. Equivalently, the greater the width W of a uniform test, the smaller is the relative standard error $SE(\beta) / SE(0)$ and the flatter the standard error curve.

Furthermore, at the extreme of the test where $\beta = \omega$,

$$SE(\omega) / SE(0) = \{ 1 + \tanh^2 (\omega/2) \}^2 < \sqrt{2} .$$

It follows that

$$SE(\omega) / SE(0) < \sqrt{2} \qquad \text{if} \qquad -\omega \leq \beta \leq \omega .$$

Thus when $-\omega \leq \beta \leq \omega$ over the test itself, the relative standard error is bounded above by $\sqrt{2}$ ensuring that the standard error curve is reasonably flat.

Flat information curves have been the aim of test designers as the result is a test instrument with constant precision over the target interval. Samejima (1983) achieved this aim with her constant information model for dichotomous data, a model with a constant information function over a finite range. Our results in this section demonstrate that this ideal can be approximated with the RDM when the test comprises a large number of items with difficulty values uniformly spread over a wide range. However, there needs to be a trade-off between test width W and test length L. For uniform tests of a given length L, the greater the width W the flatter the information curve. But increasing the test width W increases the item spacing Δ and decreases the measurement precision at the centre of the test. We address the task of balancing these conflicting requirements in later sections when we consider the task of constructing uniform tests and best test design.

To summarise the results of this section, (7) permits the calculation of the standard error of measurement at any point of the continuum of a test comprising uniformly spread dichotomous items. The standard error is a minimum at the centre of the test and increases the further the point β is away from the centre. At the centre of the test, the precision of measurement is essentially a function of the item spacing Δ only. Elsewhere on the continuum, the precision of measurement relative to the precision at the centre of the test is essentially a function of test width W only.

5. Measurement Precision with Non-uniform Tests

The theoretical ideal of a uniform spread of item difficulties can, at best, only be approximated in practice. We need to consider how to apply the results of the previous sections to tests comprising dichotomous items with a non-uniform spread of item difficulties. First, when the item difficulties are approximately uniform, the item spacings $\delta_2 - \delta_1, \delta_3 - \delta_2, \ldots, \delta_L - \delta_{L-1}$ are approximately constant and the approximations

$$\begin{array}{lll} \text{test width} & W \approx & \delta_L - \delta_1 \\ \text{item spacing} & \Delta \approx & (\delta_L - \delta_1)/(L-1) \end{array}$$

can be expected still to be valid. Substituting for W and Δ into the expressions (7) - (10), we can expect to obtain suitable results for the precision of measurement for any given test. The question then is how best to approximate the item spacing Δ, and hence the test width W, when the item spacings are no longer constant.

Evidence of non-uniformity in the distribution of item difficulties is likely to be in the extremes. Woodcock (1992) recommended using the values of the two smallest item difficulties δ_1, δ_2 and the two largest δ_{L-1}, δ_L to give

$$W \approx \left\{ (\delta_2 - \delta_1) + (\delta_{L-1} - \delta_2)(L-1)/(L-3) \right\} / 2$$

obtained by averaging the two approximations

$$\begin{array}{lll} W \approx & \delta_L - \delta_1 & \\ W = & \Delta(L-1) & \approx (\delta_{L-1} - \delta_2)(L-1)/(L-3) \end{array}$$

There are however, alternative approximations for W and Δ based on the four item difficulties $\delta_1, \delta_2, \delta_{L-1}, \delta_L$. We prefer to consider all $6 = 4(4-1)/2$ differences between pairs, namely

$$\begin{aligned}
\Delta & = \delta_2 - \delta_1 \\
\Delta\,(L-2) & = \delta_{L-1} - \delta_1 \\
\Delta\,(L-1) & = \delta_L - \delta_1 \\
\Delta\,(L-3) & = \delta_{L-1} - \delta_2 \\
\Delta\,(L-2) & = \delta_L - \delta_2 \\
\Delta & = \delta_L - \delta_{L-1}
\end{aligned}$$

Averaging these six expressions gives the approximations

$$\begin{aligned}
\Delta & \approx (3\delta_L + \delta_{L-1} - \delta_2 - 3\delta_1) / 4\,(L-2) \\
W & \approx (3\delta_L + \delta_{L-1} - \delta_2 - 3\delta_1)(L-1) / 4\,(L-2)
\end{aligned}$$

These approximations for Δ and W are very similar to Woodcock's approximations, differing only in the weightings of the item difficulties. The logical extension is to consider the values of all item difficulties $\delta_1, \delta_2, \ldots, \delta_L$ in determining an approximation for the item spacing Δ. For each pair of item difficulties δ_i, δ_j, the difference in difficulty gives the approximation

$$\Delta\,(j - i) \approx \delta_j - \delta_i \quad ; \qquad 1 \le i < j \le L.$$

Summing all $L\,(L-1)/2$ expressions gives

$$\Delta\, L\,(L^2 - 1)/6 \approx \sum_{i=1}^{L} (2i - L - 1)\,\delta_i$$

and the approximations

$$\begin{aligned}
\Delta & \approx 6 / L(L^2 - 1) \sum_{i=1}^{L} (2i - L - 1)\,\delta_i \\
W & \approx 6 / L(L+1) \sum_{i=1}^{L} (2i - L - 1)\,\delta_i
\end{aligned}$$

There is another interpretation of this latter approximation for Δ as an average item spacing. It is the slope of the least squares regression line fitting the L points with coordinates (i, δ_i); $i = 1, 2, \ldots, L$; in the Cartesian plane.

The usefulness of these approximations for the item spacing Δ and test width W depends not only on the nature and extent of the non-uniformity of the item difficulties, but also on the degree of accuracy required for the standard error. This issue is not pursued further in this article. The usefulness of expressions (7) - (10) in describing the precision of measurement of non-uniform tests would make a valuable research project.

6. Test Construction

A test is a suitable set of items specified by the set of L item difficulties $\delta_1, \delta_2, \ldots, \delta_L$ that forms the measuring instrument. Test construction involves determining an appropriate set of item difficulties that satisfies the required conditions on measurement precision over the target interval. The process of test construction, determining the test given the precision requirements, is in the opposite direction to that of understanding the precision of a given test, determining the precision properties on the target interval given the test. This relationship between precision properties and test construction is best conveyed diagrammatically as in Figure 3.

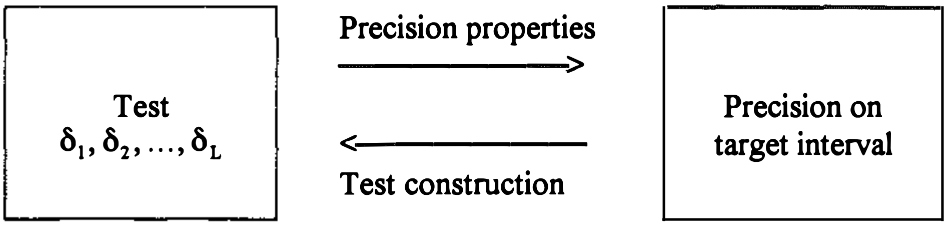


Figure 3: Relationship between precision properties and test construction

We now illustrate the use of (7) in solving problems of test construction, constructing tests with required precision requirements. The solution to each problem is best described through a numerical example. All numerical calculations required can be performed on a scientific calculator, preferably a programmable one.

Problem 1: *Construct a test given the precision at both the centre and at the extreme of the target interval.*

Example: Construct a test for which SE(0) = 0.3 and SE(2) = 0.4

Solution: From (10)

$$\mathrm{SE}(2) / \mathrm{SE}(0) = \{ (\cosh 2 + \cosh \omega) / (1 + \cosh \omega) \}^{1/2} = 0.4 / 0.3$$

Squaring,

$$(\cosh 2 + \cosh \omega) / (1 + \cosh \omega) = 16 / 9$$

and solving for cosh ω gives

$$\cosh \omega = (9 \cosh 2 - 16) / 7 = 2.551$$
$$\omega = 1.589$$

From (8), it follows that

$$\Delta = SE(0)^2 \tanh(\omega/2) = 0.059$$

Hence

$$L = 2\omega/\Delta = 2(1.589) / 0.059 = 53.42$$

and

$$W = \Delta(L - 1) = 0.059(53.42 - 1) = 3.118$$

Rounding up the test length L to L = 54, the required test has 54 dichotomous items with difficulties uniformly spaced from −1.559 logits to 1.559 logits at intervals of 0.059 logits. Note that, with L = 54, W = 3.118, SE(0) = 0.298 and SE(2) = 0.398.

Problem 2: *Construct a test given the range of item difficulties and the precision at the centre.*

Example: Construct a test for which W = 5 and SE(0) = 0.3.

Solution: As $\omega = \Delta L/2 = (W + \Delta) / 2 \approx W/2$

it follows from (8) that

$$\Delta = SE(0)^2 \tanh(\omega/2) \approx (0.3)^2 \tanh(5/4) = 0.076$$

and hence

$$L = 1 + W/\Delta \approx 1 + 5/0.076 = 66.49$$

Furthermore $\omega > W/2$, $\Delta > 0.076$ and hence $L < 66.49$. Using (8), we find when $L = 66$, that $SE(0) = 0.300$. The required test then has 66 dichotomous items with difficulties uniformly spaced from -2.5 logits to 2.5 logits at intervals of 0.077 logits.

Problem 3: *Construct a test given the range of the item difficulties and the maximum standard error over the target interval.*

Example: Construct a test for which $W = 5$ and $SE(\beta) \leq 0.5$ for $-3 \leq \beta \leq 3$.

Solution: To ensure $SE(\beta) \leq 0.5$ for all $-3 \leq \beta \leq 3$, we require $SE(3) = 0.5$. Using $\omega \approx W/2$, it follows from (7) that

$$\begin{aligned} \Delta &= SE\,(\beta)^2 \;\sinh\omega \,/\, (\cosh\beta + \cosh\omega) \\ &\approx 0.5^2 \;\sinh 2.5 \,/\, (\cosh 3 + \cosh 2.5) \\ &= 0.093 \end{aligned}$$

Hence

$$L \;=\; 1 + W/\Delta \;\approx\; 1 + 5\,/\,0.093 \;=\; 54.55$$

Using (7), we find that, when $L = 53$, $SE(3) < 0.5$ but when $L = 52$, $SE(3) > 0.5$. The required test then has 53 dichotomous items with difficulties uniformly spaced from -2.5 logits to 2.5 logits at intervals of 0.096 logits. We note further that this test has standard errors $SE(3) = 0.500$ and $SE(0) = 0.335$.

Problem 4: *Construct a test of a given length that measures as precisely as possible over a given target interval.*

Example: Given $L = 50$, $-2.5 \leq \beta \leq 2.5$, determine W that minimises the maximum standard error over $-2.5 \leq \beta \leq 2.5$, namely $SE(2.5)$.

Solution: Although we could find W by trial and error, repeatedly calculating $SE(2.5)$ for various W, it is better to proceed as follows. Substituting $2\,\omega/L$ for Δ in (7) and rearranging, we have that

$$SE(\beta)^2 \;L/2 \;=\; \omega(\sinh\omega)^{-1}\,(\cosh\beta + \cosh\omega) \;=\; F(\omega)$$

Hence $SE(\beta)$ is minimised for a given test length L and for a given β when $F(\omega)$ is minimised, namely when the derivative $F'(\omega) = 0$. Differentiating, it follows that

$$F'(\omega)\ \sinh^2 \omega \quad = \quad \sinh \omega\ (\cosh \beta + \cosh \omega) \ - \omega\ (1 \ + \ \cosh \beta \cosh \omega)$$

and hence SE(β) is a minimum when

$$\sinh \omega\ (\cosh \beta + \cosh \omega) \ -\omega\ (1 \ + \ \cosh \beta \cosh \omega) = \ G(\omega) \ = \ 0 \qquad (11)$$

For our example, $\beta = 2.5$ and $\cosh 2.5 = 6.132$. As $G(3) = -25.92$ and $H(4) = 238.74$, the value of the AHW ω that satisfies (11) and minimises SE(2.5) lies between 3 and 4. To solve for ω, we use

$$G'(\omega) \quad = \quad \sinh \omega\ (2 \sinh \omega \ - \ \omega \cosh \beta)$$

and Newton-Raphson iteration

$$\omega_{n+1} \ = \ \omega_n \ - \ G(\omega_n)/G'(\omega_n) \ ;n \ = \ 0, 1, 2, \ldots$$

as follows

$$\begin{aligned} \omega_0 &= 4 \\ \omega_1 &= 3.709 \\ \omega_2 &= 3.510 \\ \omega_3 &= 3.418 \\ \omega_4 &= 3.400 \\ \omega_5 &= 3.399 \\ \omega_6 &= 3.399 \end{aligned}$$

to obtain the solution $\omega = 3.399$. It follows that the required test width is given by

$$W \ = \ 2\omega\ (L - 1) / L \ = \ 2\ (3.399)\ 49/50 \ = \ 6.66$$

The required test, with width $W = 6.662$, has 50 dichotomous items with difficulties uniformly spaced from -3.331 logits to 3.331 logits at intervals of 0.136 logits. The maximum standard error over the target interval $-2.5 \leq \beta \leq 2.5$, namely SE(2.5) is minimised to SE(2.5) = 0.438 for all tests of 50 items. This test has standard error at the centre of the target interval SE(0) = 0.381. We note further that the 95% confidence interval $\hat{\beta} \pm 1.96\ \mathrm{SE}(\hat{\beta})$ for the person parameter β associated with a measurement $\hat{\beta} = 2.5$ is (1.64, 3.36) while that of a measurement $\hat{\beta} = 0$ is (-0.75, 0.75).

This last problem, constructing a test of a given length that measures as precisely as possible over a given target interval, may be referred to as constructing a best test. This concept of best test design will be developed further in the next section. Finally, we note that in solving problems of test construction, we have assumed that we have access to an infinitely large bank of suitable items, enabling us to select an item of any given difficulty. In practice however, an item bank is finite and we can at best only approximate any theoretical solution.

The results of this section demonstrate the advantage of uniform tests. Traditionally, a trial and error procedure (see Lord, 1980, p.72) has been used to select test items such that the resultant test has some predetermined information or standard error function. Uniform tests of dichotomous items are specified with only two test attributes, the length L and width W. Furthermore, specifying only a target interval rather than a target distribution of person abilities β, simplifies the concept of a target to the extent that not only can problems of test construction be well defined, they are sufficiently tractable to be readily solved.

7. Best Uniform Test

The solution to problem 4 in the previous section illustrates the construction of a best uniform test, the test of a given length that measures as precisely as possible over a given target interval. It satisfies the criterion for a best test described by Wright and Stone (1977) as follows.

> A best test is one which measures best in the region within which measurements are expected to occur. Measuring best means measuring most precisely. (p.133)

The region within which measurements are expected to occur is the target interval $B = [-\beta_0, \beta_0]$ specified by the extreme value β_0 of the parameter β. How well any given test measures in the region is specified by the least precise measurement in the interval, namely

$$\max_{\beta \text{ in } B} \mathrm{SE}(\beta)$$

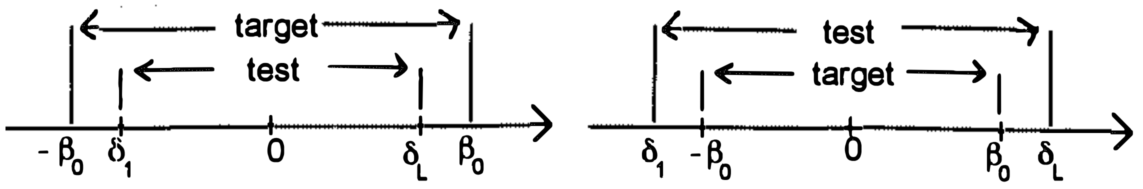
which for uniform tests, is given by the standard error $\mathrm{SE}(\beta_0)$ at the boundary of the interval.

Note that we make no assumption about the distribution of target persons, only the boundary value of the interval in which they are expected to be located. The best uniform test, the one that measures most precisely, is the one that minimises this maximum standard error. Now (7) shows that, for a given test width W, $SE(\beta_0)$ can be made arbitrarily small by making the test length L arbitrarily large. In practice, we make a test as long as is feasible within the constraints of available time and the attention span of the target persons. Hence the length L of a best test needs to be specified. Finding the best uniform test is then finding the test width W, or equivalently the adjusted half width ω, that minimises this maximum standard error as follows.

$$\min_{\omega} \quad \max_{\beta \text{ in } B} \quad SE(\beta) = \min_{\omega} SE(\beta_0)$$

Thus the best uniform test satisfies a 'minimax condition' on the standard error of measurement. It is the best uniform test of a given length on a given interval. Specifying only a target interval B rather than a target distribution of person abilities and restricting the possible tests to those with a uniform distribution of item difficulties, permits a ready solution to the problem of constructing a best test. Following the procedure illustrated in problem 4, for any given target interval B and test length L, we can always determine the test width W and so construct the best uniform test of length L on the interval B. We now describe further properties of best uniform tests.

First, we note that the width W of the best uniform test need not coincide with the width of the target interval B. For example, the solution to problem 4 in the previous section considers a target interval $[-2.5, 2.5]$ of width 5, and shows that the best uniform test of length $L = 50$ has width $W = 6.662$. Thus the extreme item difficulties δ_1 and δ_{50} lie outside the target interval. For some examples, the item difficulties of the best uniform test may be expected to lie entirely within the target interval, for others the extreme item difficulties may be expected to lie outside the target interval. The relationships between the best test and target interval is illustrated in Figure 4.



(i)

test within target $(\beta\ > \delta\ )$

(ii)

target within test $(\beta_0 < \delta_L)$

Figure 4: Relationship between the target interval and the best test

This refines the conclusion of Wright (in the afterword of Rasch, 1960,)

> a uniform distribution of item difficulty from one end of the target to the other, approximates the best possible test design in most real situations. (p.194)

that the intervals will coincide.

Rarely will the extremes of the target interval and those of the item difficulties of the best uniform test coincide. We now obtain the relationship between the boundary value β_0 of the target interval $B = [\ -\beta_0, \beta_0\ ]$ and the adjusted half width ω of the best uniform test on B of a given length L. From (11) it follows that

$$\cosh \beta_0 \quad = \quad (\sinh \omega \cosh \omega - \omega) / (\omega \cosh \omega - \sinh \omega) \qquad (12)$$

Thus the AHW ω of the best uniform test depends only on the target interval B and not on the length L. Equivalently, all best uniform tests of different lengths L on the same target interval B have the same value of the AHW ω.

Consider now the solution defined by (12) for narrow best tests with arbitrary small values of the AHW ω. As $\omega \rightarrow 0$, the value of the right hand side of (12) and hence that of $\cosh \beta_0 \rightarrow 2$. It follows that (12) provides a solution to the best uniform test provided that $\cosh \beta_0 > 2$, or equivalently, $\beta_0 > \cosh^{-1}2 = 1.317$. If the target interval B has boundary value $\beta_0 \leq \cosh^{-1}2$, the best uniform test obtained by minimising SE(β) on B, has AHW $\omega = 0$. Hence, $\omega = \Delta = W = 0$ and, irrespective of test length L, the best uniform test on B has all L items of equal difficulty, $\delta_1 - \delta_2 - \ldots - \delta_L - 0$, located at the centre of the target interval.

The boundary value β_0 of the target interval $B = [-\beta_0, \beta_0]$ is equal to the AHW ω of the best uniform test when

$$\cosh \omega = (\sinh \omega \cosh \omega - \omega) / (\omega \cosh \omega - \sinh \omega)$$

namely $\beta_0 = \omega = 1.606$. Now from (12), cosh β_0 is an increasing function of ω and hence ω is an increasing function of cosh β_0 and therefore of β_0, $\beta_0 > 0$. It follows that, if $\cosh^{-1}2 = 1.317 \leq \beta_0 < 1.606$, $\omega < \beta_0$ and the best uniform test lies entirely inside the target interval. For wider target intervals with $\beta_0 > 1.606$, $\omega > \beta_0$ and the extreme item difficulties of the best uniform test will lie outside the target interval.

To summarise, specifying a distribution-free target interval B rather than a target distribution of person abilities β has simplified the concept of a target. Restricting our consideration of tests to those with a uniform distribution of item difficulties has led to a concept of a best uniform test as the test that satisfies a 'minimax condition' on the standard error of measurement. This concept leads to a unique best uniform test that can readily be found in practice.

8. Summary

This article has examined the precision of measurements obtained from using the Rasch Dichotomous Model to analyse test data. Considering tests in which the item difficulties are uniformly spaced from easiest to most difficult, permits the derivation of an alternative expression for the standard error of measurement. This expression is sufficiently simple to enable the precision properties of uniform tests to be readily described and to enable a variety of problems of test construction to be solved. One particular problem is that of best test design. Regarding measurement precision as a property of the test only, we have been led to consider a 'distribution-free' target interval. The best uniform test of a given length on a given target is the one that satisfies a minimax condition on the standard error. We illustrated the solution to this problem and described properties of best tests.

However, there is not much comfort to be gained from this best test design. Our example in Section Four determined the best uniform test of 50 items on the target interval $[-2.5, 2.5]$. That the minimum width for a 95% confidence interval for a measurement $\hat{\beta} = 25$ at the end of the interval is as large as $3.36 - 1.64 = 1.72$ logits should be a matter of concern for researchers who wish to obtain precise measurements. This limitation in precision arises not from the Rasch Dichotomous Model as such, but rather the dichotomous 0/1 format of item responses.

Samejima (1969 has shown that the use of more than two ordered categories for item responses provides more information about the value of each person parameter than dichotomously scored items. This result suggests that the use of more than two ordered categories for item responses should result in greater precision of measurement for a test of the same length. Item response models for the analysis of partial credit data have been developed. (De Ayala, 1993) but properties of measurement precision arising from the use of such models are not yet understood. Extending the methods developed in this article from the Rasch Dichotomous Model to such partial credit models should not only assist in understanding their properties, but also show the way to researchers seeking to construct tests with greater measurement precision.

9. References

Anderson, E.B. (1980). *Discrete statistical models with social science applications.* Amsterdam: North Holland.

De Ayala, R.J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counselling and Development*, 25, 172-189.

Douglas, G.A. (1974). *Test design strategies for the Rasch psychometric model.* Unpublished doctoral dissertation, University of Chicago.

Haberman, S.J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics,* 5, 815-841.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* New Jersey: Lawrence Erlbaum.

Rao, C.R. (1973) *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedoigogiske Institute, 1960. Reprinted Chicago: University of Chicago Press.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Pychometrika* Monograph Supplement No. 17.

Samejima, F. (1983). The constant information model on the dichotomous response level. In D.J. Weiss (Ed.), *New horizons in testing* (pp. 287-308). New York: Academic Press.

Swaminathan, H. Parameter estimation in item response models. In R.K. Hambleton (Ed.), *Applications of item response theory.* Vancouver: Educational Research Institute of British Columbia.

Woodcock, R. (1992). Woodcock's nomograph. *Rasch Measurement*, 6, (3), 243-244.

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