Measurement and Research Department Reports 92-1

A Logistic Model For Time Limit Tests

N.D. Verhelst
H.H.F.M. Verstralen

M.G.H. Jansen
University of Groningen

Cito

Measurement and Research Department Reports 92-1

A Logistic Model For Time Limit Tests

N.D. Verhelst
H.H.F.M. Verstralen

M.G.H. Jansen
University of Groningen

Cito
Arnhem, 1992

Abstract

Using a result by Dubey (1969) a logistic model for time limit tests is developed. In its present form all items are assumed to have equal discriminatory power. The likelihood of this model, however, resists simplification towards a form appropriate for calculations, in particular it does not yield the attractive result of Dubey if more than one item is involved. The main obstacle is an interdependence of conditional times spend on the items, given that n items are finished in the allotted time. Therefore, an approximate pseudolikelihood function is proposed that is proven to be asymptotically equal to a pseudolikelihood. Using a result from Arnold and Strauss (1988), maximization of this function yields consistent estimates of the model parameters. The pseudolikelihood circumvents the interdependence, and the approximation returns Dubeys result. With this function the derivation is straightforward and presents no difficulties. Available programs for the estimation and testing of item and person parameters in logistic models are applicable. The estimation of the subject parameters is evaluated for realistic testing situations by a series of simulations.

Introduction

In the influential textbook by Lord & Novick (1968) we find the warning that test and measurement do not have identical meanings, although in general these terms are used interchangeably. Their warning concerns the need for distinction in relation to the differences between speed and power tests. If we take a person's total incorrect score, which is the number of items she has not answered correctly, then it is clear that this score may be viewed as the sum of items to which the subject gave the wrong answer and of items that the subject did not attempt. Lord and Novick distinguish two kinds of measurements. If all subjects are given sufficient time to complete all the items, and in fact do so, then the total incorrect score will consist of wrong answers only. A test administered in this way will be called a power test. At the other extreme, suppose that all items of the test are such that every subject who attempts the item answers it correctly. Assume further that there are so many items that no subject will be able to complete all items, within the time-limits specified for the test. Since subjects differ in speed, they will tend to complete a different number of items in any fixed time period. Such a test is called a pure speed test. Note that whether or not a test is a power test does not depend on the content only but also on the conditions under which it is administered. Tests employed in practice are rarely pure speed or pure power tests; most tests are what we call partially speeded, with the degree of speededness depending on the actual time-limits. Already in the early classical test theory literature it was recognized (Gulliksen, 1950) that speededness had to be taken into account. Attempts have been made to develop indexes of speededness (Cronbach & Warrington, 1951). Some reliability theory for speeded tests has been given by Guttman (1955).

Another line of test research on time-limit tests concerns the use of Poisson process models to develop strong true score models for the responses on highly speeded tests (Rasch, 1960; Van der Ven, 1969). Meredith for instance (Meredith, 1971) develops an axiomatic model for tests of pure speed, leading to the conclusion that such test scores may, under certain circumstances, be regarded as realizations of a poisson process and Moore (Moore, 1970) applies poisson process models to six highly speeded tests from the Kit of Reference Tests for Cognitive Factors with varying success. Models for mental test scores resulting from time-limit tests that take a combination of speed and power effects into account are described by among others, Iseler (1970), White (1973), Pieters and Van der Ven (1982). Pieters and Van der Ven have shown that the Rasch model and the binomial error model can be used to estimate precision, focusing on the number of items correct and the number of items attempted. To estimate speed, a model called the Poisson-Erlang model is presented.

Recently, certain modifications of the one-parameter logistic Rasch model, which are suited to modeling time-limit tests have been introduced (Roskam, 1987; Van Breukelen, 1989). Our model is superficially similar to Roskam's model, which was inspired by a proposal of Scheiblechner. For instance, in Roskam's model it is assumed that the probability of giving the correct answer to an item depends on the difficulty of the item and the 'effective' ability of the subject, given a certain response time. The effective ability increases, linearly for example, with the amount of time invested. According to this model the probability of a correct response increases with subject ability and response time and decreases with item difficulty. Roskam also included a probability distribution for the item response times. Van Breukelen who showed that the parametrization gave rise to some implausible interpretations formulated a modification, where Roskam's Rasch model for the response probabilities was combined with a Weibull distribution for the response times. This last model contains an extra subject parameter, representing the subjects's willingness to invest time in solving an item. This subject parameter influences the probability of a correct answer indirectly by influencing the response times. A typical feature of this and some of the earlier models is the assumption, based on arguments derived from psychological theory concerning the item solving process, that there is some kind of intra-subject trade-off process between speed and power. Estimating the parameters of the Rasch-Weibull model proved to be problematic as was also shown by Donders (1990).

With the kind of time limited tests, we try to model, students have a fixed amount of time τ, for example let $\tau = 30$ minutes, to take the test. They are encouraged to respond correctly with a reasonable certainty to as many items as they can in time τ. It is not allowed to skip items. A skipped item counts as a wrong response. It is assumed that no student finishes the whole test within the allotted time. The model is developed for a time limited test with binary scored items, where the observations are as follows: for every student v, his response pattern $\underline{x}_v$ and the number n_v of finished items in time τ are observed. No data on the individual response times per item are available. The only information on the response times is the number n_v of finished items in time τ.

The three main concepts in the model related to a person are speed, precision, and mental speed.

Speed is related to the time a student works on a problem until he gives a response. The faster he is the shorter the time per item. It is to be stressed that in the model speed is only a time parameter.

Precision is related to the conditional probability of a correct response, given that he responds to an item. The greater the precision of a student the greater the probability that he responds correctly to an arbitrary item.

Mental speed, further called fluidity to avoid confusion with speed, can be understood as the amount of adaptive mental change a person can realize per unit time. Adaptive mental change is supposed to cause the probability of a correct response to increase.

The main concept related to an item is its difficulty. Difficulty in this context shows a little shift in meaning with respect to the more familiar IRT models. Here difficulty is related to the amount of adaptive mental change the item needs to realize a certain increase in the probability of a correct response.

The primary aim of the model is to estimate the fluidity of a student as a function of his speed and precision.

The following considerations guided the choice of the model:

1. The model should allow for the estimation of a person parameter ν_v, the fluidity of person v, from his speed and accuracy. In order to do this properly, the model should specify how the amount of time that a certain fluidity is applied to an item with a certain difficulty affects the probability of a correct response: the precision.

2. Response times are allowed to show variation within subjects. So, part of the model will be a response time distribution characterized by a subject speed parameter, β_v, v denoting the v-th subject. In a population under research β can be considered a random variable, with a certain distribution. It is assumed that the parameters of this distribution, in particular the location parameter, are a function of the instructions and the general setting under which the data are collected. We propose not to impose a specific structure on this distribution, as an integral part of the model. In the same vein it is possible to conceive of the latent variable ν, measured by the test items, as a variable whose distribution is not considered as part of the measurement model.

3. Items may differ in difficulty, denoted by ρ. Items can differ in the amount of adaptive mental change needed to achieve a certain probability of a correct response.

4. Separation of fluidity, speed and item parameters. The joint distribution of (ν,β) conditional on ρ could, in principle, be chosen such that a stochastic dependence between fluidity, speed and item difficulty is modeled. Our viewpoint, however, is that this dependence can be quite complicated. Therefore, the dependence between ν, β, and ρ is preferably investigated empirically, and not formally incorporated in the model. However, the dependencies between these three parameters do not affect the model in the same way. Even a strong dependence between ν and β will not violate the

model assumptions, a dependence between ρ of an item and the time devoted to it, however, will.

This approach puts restrictions on the model, because it implies that the item parameters can be estimated from the model independently of the dependence between ν and β. As will become clear in the sequel, the main restriction amounts to assuming independence of the response time distribution from the item parameters. The time devoted to an item, therefore, is assumed to be independent of the difficulty of the item. It only depends on the speed parameter of the student. This assumption can probably only be upheld if the items do not show large differences in difficulty. In that case a trade-off between invested time and precision may result in essentially longer invested time for difficult items.

This approach may be contrasted with the model developed by Van Breukelen (1989), where the parameter of the response time distribution (for each item) is a specified function of item difficulty, fluidity and a specific (subject)parameter for speed, called persistence.

5. A logistic marginal model with time integrated out. As already mentioned, the time per item is not observed.

The Model

The usual notational conventions are adopted: Random variables are denoted by Roman capitals, a realization by their lower case equivalent. Likewise, distribution functions are denoted by Roman capitals, their densities by the corresponding lower case.

The choice of model distributions is guided by a result due to Dubey (1969), see also (Johnson & Kotz, 1970, p 289), who showed that the compound distribution of a gamma distribution and a generalized exponential distribution gives rise to a generalized logistic distribution. The latter distribution has an obvious relation with Rasch models, where we find ourselves on better known territory.

The time spent on an item is represented by the random variable T assumed to be distributed as:

$$\frac{d}{dt} P(T \leq t) = g(t) = \frac{\beta^p}{\Gamma(p)} t^{p-1} e^{-\beta t}, \qquad (t \geq 0;\ p > 0;\ \beta > 0). \tag{1}$$

(1) is called the gamma distribution, where β/p here represents the expected number of finished items per unit time, and, therefore, may be interpreted as a speed parameter.

Let α be a general scale parameter. Further let

$$y_{vi}(t) = \alpha t \frac{\nu_v}{\rho_i} \quad , \tag{2}$$

then the conditional probability of a correct response by v on i given the time T=t is given by:

$$P(+;\nu,\rho,\alpha \mid T = t) = H_t(\nu,\rho,\alpha) = 1 - \exp[-y_{vi}(t)] \quad , \tag{3}$$

an exponential distribution. The role and meaning of α will be discussed below.
Denote the expected time per item using the distribution given by (1) with $\bar{T}_\beta = p/\beta$. Although he used the exponential parametrization given in the next section, Dubey (1969) shows that the marginal probability of a correct response with time t integrated out is then given by:

$$\begin{aligned} P(+;\nu,\beta,\rho,\alpha) &= \int_0^\infty g(t) H_t(\nu,\rho,\alpha) dt \\ &= 1 - \left[1 + \alpha \frac{\bar{T}_\beta}{p} \frac{\nu}{\rho}\right]^{-p} , \end{aligned} \tag{4}$$

a generalized logistic distribution. The familiar logistic distribution is obtained from (4) by setting p=1. In that case (1) simplifies to the exponential distribution:

$$P(T \leq t) = G(t) = 1 - \exp(-\beta t), \quad (t \geq 0; \beta > 0), \tag{5}$$

and accordingly β denotes the number of items finished per unit time. In this simple case the first part of (4) can be considered a Laplace transform of H and the result follows easily. It should be noted that the exponential distribution has constant hazard rate β. Therefore, the conditional probability distribution to stop with an item at $T - t_0$ ($T > t_0$), given that the student is still working on it on time t_0 is identical for every instant t_0, that is, it is independent of the time already spent on the item. Although a constant hazard rate may not seem realistic, (5) will be used as response time distribution function further on. By choosing the scale parameter p in (1) different from 1, a hazard rate which is dependent on t is obtained. The estimation problems, however, become more complicated, even in the case p

is fixed in advance. Detailed derivations of the more general case will be described in a future report.

Using (5) and (2), the marginal probability (4) that student v with ability ν_v and speed β_v responds correctly on item i with parameter ρ_i can be simplified to:

$$P(+;\nu,\beta,\rho,\alpha) = \frac{y_{vi}(\bar{T}_\beta)}{1 + y_{vi}(T_\beta)}. \tag{6}$$

If y is interpreted as a measure of 'Readiness' to solve the item connected to the probability of a correct response by (3) or its marginal (6), then (2) clearly shows that ν is a fluidity or mental speed parameter. To clarify this assume that, at the start of attempting an item the state of a persons mind is reset to 'neutral' and that the time trying to solve it is interpreted as a movement with constant speed ν/ρ along the axis that measures readiness represented by y. The time t times ν/ρ gives the amount of readiness y reached after time t. The item parameter ρ represents the amount of mental change needed to arrive at a unit of readiness to solve the item.

The meaning of α can now be elucidated as well. If the unit of mental change is arbitrary then α is arbitrary. It can be chosen to be equal to 1, and, therefore be omitted. If, however, the unit of mental change is not arbitrary, for instance by fixing the variance of fluidity in a population to unity (given a time unit), or by a cognitive theory, then α can and must be estimated.

A nice interpretation of the function of the probability of a correct response is arrived at by introducing the following reparametrisation: $\nu = \exp(\xi/\alpha)$ and $\rho = \exp(\epsilon/\alpha)$. The exponential parameters ξ and ϵ will also be called fluidity and difficulty respectively, because it is clear from the context what is meant. Now assume a random variable Z that represents the 'temporary fluidity' of a person. Formula (3) can now be written as a generalization of the double exponential distribution, as originally used by Dubey (1969):

$$P(Z \leq \varepsilon;\xi,\alpha \mid T=t) = \exp(-t\alpha \exp[(\xi-\varepsilon)/\alpha]). \tag{7}$$

By writing (7) in the form:

$$P(Z \leq \varepsilon \mid t) = H_t(\varepsilon;\xi,\alpha) = \exp[-\exp\{([\xi + \alpha \ln t\,\alpha]-\varepsilon)/\alpha\}],$$

it is easily recognized that the time spent results in a shift $\alpha \ln t\alpha$ of the location of the temporary fluidity distribution.

If, analogous to Thurstone's (1927) law of comparative judgement, the probability of a correct response on item i is identified with the probability that the temporary ability Z exceeds the value of the item parameter ϵ_i we rewrite (4) as:

$$P_i(+ \mid \xi) = \int_0^{\infty} \int_{\varepsilon_i}^{\infty} g(t)\, h_t(z)\, dz\, dt$$

$$= 1 - [1 + \alpha\beta^{-1} \exp((\xi - \varepsilon_i)/\alpha)]^{-p} .$$

Expressions equivalent to (2) to be substituted in (6) are:

$$y_{vi}(\bar{T}_{\beta_v}) = \alpha \bar{T}_{\beta_v} \exp[(\xi_v - \varepsilon_i)/\alpha]$$

$$= \exp\left[\frac{\xi_v - \varepsilon_i}{\alpha} + \ln \alpha \bar{T}_{\beta_v}\right]$$

$$= \exp\left[\frac{\xi_v}{\alpha} - \left[\frac{\varepsilon_i}{\alpha} - \ln \alpha \bar{T}_{\beta_v}\right]\right] \qquad (8)$$

$$= \exp\left[\left(\frac{\xi_v}{\alpha} - \ln \frac{\beta_v}{\alpha}\right) - \frac{\varepsilon_i}{\alpha}\right] .$$

The expression within parentheses in the fourth part of (8) can be considered the precision parameter θ of the person, and is comparable to the person parameter in the Rasch model. This shows that in the exponential parametrisation we have:

$$\frac{\xi}{\alpha} = \theta + \ln \frac{\beta}{\alpha} \qquad (9)$$

fluidity = precision + speed

The expression within parenthesis in the third part of (8) shows an aspect of the model that raises some caution with respect to item parameter estimates in relation to the model assumption that time spent on an item is not related to its difficulty. Suppose, however, that, contrary to the model assumption, there exists a positive relation between difficulty and invested time, then the parameter of a difficult item will be underestimated, and overestimated for an easy item, or, what amounts to the same, the variance of the item

difficulties will be underestimated. Probably, therefore, if the variance is not too large, the neglect of a possible relation between time and difficulty will not cause serious problems.

We cherish this assumption because it yields a very precious property of the model: a sufficient statistic for β (see the next section).

Maximum Likelihood Estimation

Assume that the test is administered to V students, and let N_v, with realizations n_v, denote the number of finished items in time τ by student v (v = 1, ..., V). Let the response variable be denoted by $\underline{X}_v = (X_{v1}, X_{v2}, .., X_{vn_v})$,with $X_{vi} = 1$ for a correct answer, and $X_{vi} = 0$ for a wrong answer. The basic observations under the model consist of the pair $(n_v, \underline{x}_v)$ and the likelihood for one observation is proportional to the joint probability of this pair. Denote the set of observed response patterns by $\{\underline{x}\}$ and the set of 'items finished by the respondents' by $\{n\}$ and let $\underline{\xi} = (\xi_1, \ldots, \xi_V)$, $\underline{\beta} = (\beta_1, \ldots, \beta_V)$ and $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)$, with $m = \max_v(n_v)$. Denote the complete set of observations with $\{\underline{x}, n\}$.

The exponential distribution has the following useful property for the estimation of β. Let $T_1, T_2, \ldots$ be a sequence of independent random variables identically distributed as the exponential distribution with parameter β. Let τ be a fixed time interval, then the random variable N defined by:

$$\sum_{i=1}^{N} T_i \leq \tau < \sum_{i=1}^{N+1} T_i ,$$

is Poisson distributed with parameter $\tau\beta$. The interpretation for our test situation is straightforward. If the limited test time equals τ and a student v has speed parameter β_v then his number N_v of finished items has the indicated Poisson distribution. Moreover, the following derivation shows that the conditional density $f(\underline{t};\beta \mid n) = f(\underline{t} \mid n)$ is independent of β.

$$f(\underline{t}; \beta \mid n) = \frac{\beta^n \exp(-\beta \sum_i^n t_i) \exp[-\beta(\tau - \sum_i^n t_i)]}{\frac{(\tau\beta)^n \exp(-\tau\beta)}{n!}} = f(\underline{t} \mid n).$$

Therefore, n is sufficient for β and its unbiased maximum likelihood estimator is n/τ.

The likelihood of the complete data now factors in:

$$L(\underline{\xi}, \underline{\beta}, \underline{\varepsilon}, \alpha; \{\underline{x}, n\}) = L_P(\underline{\beta}; \{n\}) \times L_C(\underline{\xi}, \underline{\varepsilon}, \alpha; \{\underline{x}\} \mid \{n\}) \ ,$$

with L_P the above mentioned Poisson probability function, and the conditional probability function L_C, using (3):

$$L_C(\underline{\xi}, \underline{\varepsilon}, \alpha; \{\underline{x}\} \mid \{n\}) =$$

$$\prod_v^V f(\underline{t} \mid n_v) \int_{\Delta_n} \prod_i^{n_v} [1 - \exp(-y_{vi}(t))]^{x_{vi}} \exp(-y_{vi}(t))^{1-x_{vi}} (dt)^n \ , \qquad (10)$$

with $\Delta_{n,\tau}$ the set of n-dimensional vectors $\{\underline{t} : \Sigma t_i \leq \tau, t_i \geq 0\}$. We omit the subscripts of Δ, n or τ, if they are clear from the context.

Now consider the conditional likelihood L_C in (10). Because $f(\underline{t}|n)$ is constant over Δ, it can be discarded, because the likelihood is invariant under multiplication by a constant. However, this expression of the likelihood resists further simplification. The main obstacle is the constant density of $\underline{t}$ over Δ, which introduces an interdependency among the individual components t_i ($i = 1, \ldots, n_v$) of $\underline{t}$, due to the inequality constraint $\sum t_i \leq \tau$. Maximizing (10) with respect to $\underline{\xi}$ and $\underline{\varepsilon}$ therefore seems prohibitive. Moreover, it is not desirable either: the number of elements in $\underline{\xi}$ grows at the same rate as the sample size, and therefore it is to be expected that consistency of the ML-estimators does not hold, not only for $\underline{\xi}$ but also for $\underline{\varepsilon}$. So, maximization of the likelihood function (10) is abandoned as an estimation procedure.

Pseudolikelihood Estimation

Arnold and Strauss (1988) showed that consistent estimates of model parameters obtain if a product of marginal and/or conditional likelihoods is maximized instead of the likelihood function itself. In the present model we replace the inner integral in L_C by a product of marginal likelihoods. For each item response we obtain the marginal density from the joint density of the response pattern of a person. A simple example with only two items 1 and 2 finished by a person which may clarify the procedure is given in Appendix A.
From Appendix A it is not difficult to verify that, in general, with n items finished we get:

$$PM_i(x_i|n) = \frac{n!}{\tau^n} \int_0^\tau \left[1 - \exp\left[-\alpha t \frac{\nu}{\rho_i}\right]\right]^{x_i} \exp\left[-\alpha t \frac{\nu}{\rho_i}\right]^{1-x_i} dP_{\Delta_n}(T \leq t) \ ,$$

where

$$P_{\Delta_n}(T \leq t) = 1 - \left(1 - \frac{\frac{n}{\tau}t}{n}\right)^n .$$

The conditional pseudolikelihood function PL_C now is given by:

$$PL_C(\xi, \underline{\varepsilon}, \alpha; \{\underline{x}\} \mid \{n\}) = \prod_v^V \prod_i^{n_v} PM_i(x_{vi} \mid n_v). \tag{11}$$

$PM_i(x_i)$ still is not equal to the logistic model. Fortunately, however,

$$\begin{aligned} P_{\Delta_n}(T \leq t) &= 1 - \left(1 - \frac{\frac{n}{\tau}t}{n}\right)^n \\ &\approx 1 - \exp(-mt) \end{aligned} \tag{12}$$

$$\text{with } \hat{\beta} = \frac{n}{\tau} \leq m \leq \frac{n+1}{\tau} .$$

Appendix B gives some considerations to select a proper value for m from this interval.

Using the approximation (12), and using (2), (4) and (6), the conditional pseudolikelihood (11) is replaced by the approximate conditional pseudolikelihood:

$$@PL_C(\xi, \underline{\varepsilon}, \alpha; \{\underline{x}\} \mid \{n\}) = \prod_v^V \prod_i^{n_v} \frac{y_{vi}\left(\frac{1}{m}\right)^{x_{vi}}}{1 + y_{vi}\left(\frac{1}{m}\right)} . \tag{13}$$

For $@PL_C$ to approximate PL_C it should also be noted that $H_t(.)$ is bounded within the real interval (0,1).

Now, for the derivation of the estimation equations let u_i be the item-score:

$$u_i = \sum_{\substack{v \\ n_v \geq i}} x_{vi} .$$

And s_v the raw sore of v:

$$s_v = \sum_i^{n_v} x_{vi} ,$$

and denote max $\{n_v\}$ with m. Further let θ_v be defined by (9) and define

$$\sigma_i = \frac{\varepsilon_i}{\alpha} ,$$

then (13) can be written as

$$@PL_C(\xi, \underline{\varepsilon}, \alpha; \{\underline{x}\} \mid \{n\}) = \frac{\prod_i^m \exp(-\sigma_i u_i) \prod_v^V \exp(\theta_v s_v)}{\prod_v^V \prod_i^{n_v} [1 + \exp(\theta_v - \sigma_i)]} . \qquad (14)$$

Clearly, (14) is an exponential family likelihood, readily recognized as a slightly generalized Rasch model. By conditioning on the sufficient statistics s_v (14) can be rewritten as

$$@PL_C(\xi, \underline{\varepsilon}, \alpha; \{\underline{x}\} \mid \{n\}) = \frac{\prod_v^V \prod_i^{n_v} \exp(-\sigma_i x_{vi})}{\prod_v^V \gamma_{s_v}(e^{-\sigma})} \times \frac{\prod_v^V \gamma_{s_v}(e^{-\sigma}) \exp(\vartheta_v s_v)}{\prod_v^V \prod_i^{n_v} [1 + \exp(\vartheta_v - \sigma_i)]} \qquad (15)$$

$$= M_C(\underline{\varepsilon}, \alpha; \{\underline{x}\} \mid \{s\}) \quad \times \quad M_M(\xi, \underline{\varepsilon}, \alpha; \{s\}) ,$$

where $\gamma_s(.)$ represents the well known basic symmetric function of order s, and the argument $e^{-\sigma}$ denotes $(\exp(-\sigma_1), \ldots, \exp(-\sigma_{n_v}))$. So the item parameters may be estimated by maximizing the first factor in (15), which is nothing else than CML-estimation in an incomplete design.

In order to estimate the subject parameters ξ, it is common use among the practitioners of the Rasch model to maximize the likelihood function with respect to the subject parameters, while keeping the item-parameters fixed at their CML-estimates. This amounts to a procedure, known as restricted maximum likelihood (REML, Amemiya, 1986), which, in general, does not affect the consistency of the estimates. If we proceed in the same way, the function to be maximized is

$$s_v = \sum_i^{n_v} P_{vi}(+; \theta \mid \hat{\varepsilon}_i, \alpha) \qquad (v = 1, \ldots, V).$$

For θ, and then using the estimate for β and (9) an estimate for ξ is obtained.

Bias in Ability Estimation

It is well known (Lord, 1983) that maximum likelihood estimates of the subject parameters in IRT-models are biased, in the sense that extreme scores get too extreme parameter estimates. Warm (1989) proposed an estimation procedure which reduces the bias up to the order 1/n, called weighted maximum likelihood (WML). Of course, in the present model one could estimate ϑ by this procedure, instead of maximizing the second part M_M of the approximate pseudolikelihood (15). Since the estimate of β is unbiased, one might hope that the resulting estimate of ξ will be unbiased, at least up to order 1/n.

In a small simulation study, a test consisting of 50 items was used. The item parameters ϵ of the first 5 items were -2, -1, 0, 1 and 2 respectively. This sequence was repeated ten times, so that at whatever point test taking is finished, reasonable information is obtained over a broad range of the latent variable. The study comprised two facets: (1) 21 different values of ξ were used equally spaced in the interval [-2, 2]; (2) number of items (n) finished. This number took the values 5(5)50. So for every combination, we want to estimate the bias defined as

$$\text{bias}(\xi, n) = \xi - E(\hat{\xi} \mid \xi, n). \tag{16}$$

The time limit of the test (τ) is 30 (unspecified) time units. Since the bias, conditional on the number of finished items is investigated, the value of β is irrelevant; α was set to 1. Generation of artificial data following the model for a combination of ξ and n was accomplished as follows:

Step 1. From the uniform distribution on $\Delta = \{\underline{t} \mid t_i > 0, \Sigma t_i \leq 30\}$, one $\underline{t}$ was sampled. The number of elements in $\underline{t}$ is n. The technical details of this sampling are explained in Appendix C.

Step 2. Using each component of the sampled $\underline{t}$, the probability of a wrong answer on each of the n finished items is computed, using (3) or (7).

Step 3. For each of the n items, an independent uniformly distributed random number is generated (in the (0,1) interval. If this number is larger than the probability calculated in step 2, the response is considered as correct.

Step 4. The number of correct responses is used to compute the WML-estimate of ϑ. The estimate of ξ is given by using $\hat{\beta} = n/\tau$ and (9).

Steps 1 to 4 were repeated 3000 times for each (ξ, n) combination. The average of the 3000 estimates was used as an estimate of the expected value in (16).

The results are displayed graphically in Figure 1. In the first part, it is seen that there is considerable bias for combinations of large ξ and small n. Small n means that the average time for answering an item is large, and, as was explained earlier, a large response time acts as a positive shift of the location parameter of the extreme value distribution. So in those cases, the test is relatively easy, so that many high scores will occur. It is well known (see Warm, 1989 and Verhelst & Kamphuis, 1990) that WML-estimates are positively biased. The second plot of Figure 1 displays the same information, but with the n-axis reflected. It is clear that for most cells the bias is very low, and has a uniform appearance. In order to display this uniformity, the third part displays the bias for values of n ranging from 15 to 50,in order to enlarge the resolution of the plot for the larger n. The pattern of the surface is generally quite irregular, the irregularities being caused by sampling error.

It seems that the estimation method works quite well, as long as the number of items finished is not too small. Of course, the design of the test was advantageous to this result. Presenting the items in increasing order of difficulty makes that the average difficulty of the finished items increases systematically with n: a considerable bias for moderate and large ξ-values results, even for moderate n, see Figure 2. Note that the values at the z-axis (Bias) are larger in Figure 2 compared to the corresponding plots in Figure 1.

Bias as a function of ξ and n

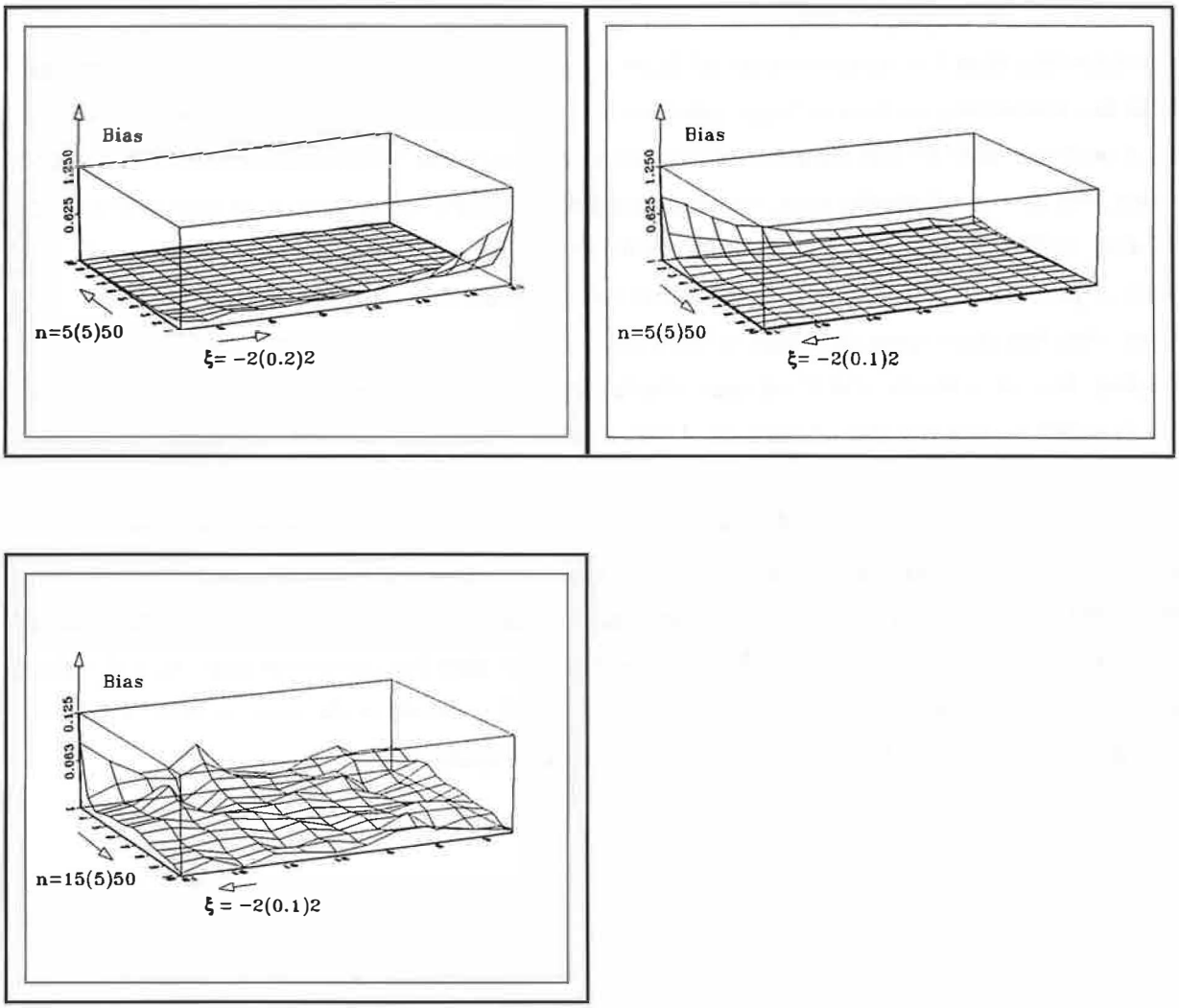

Figure 1: In the first plot ξ and n increase to the right and to the back, as usual. The second plot is the first viewed from the backside. The third plot is the second restricted to n = 15(5)50 to enlarge the resolution of the plot for larger n.

Bias as a function of ξ and n
with item difficulty increasing with position

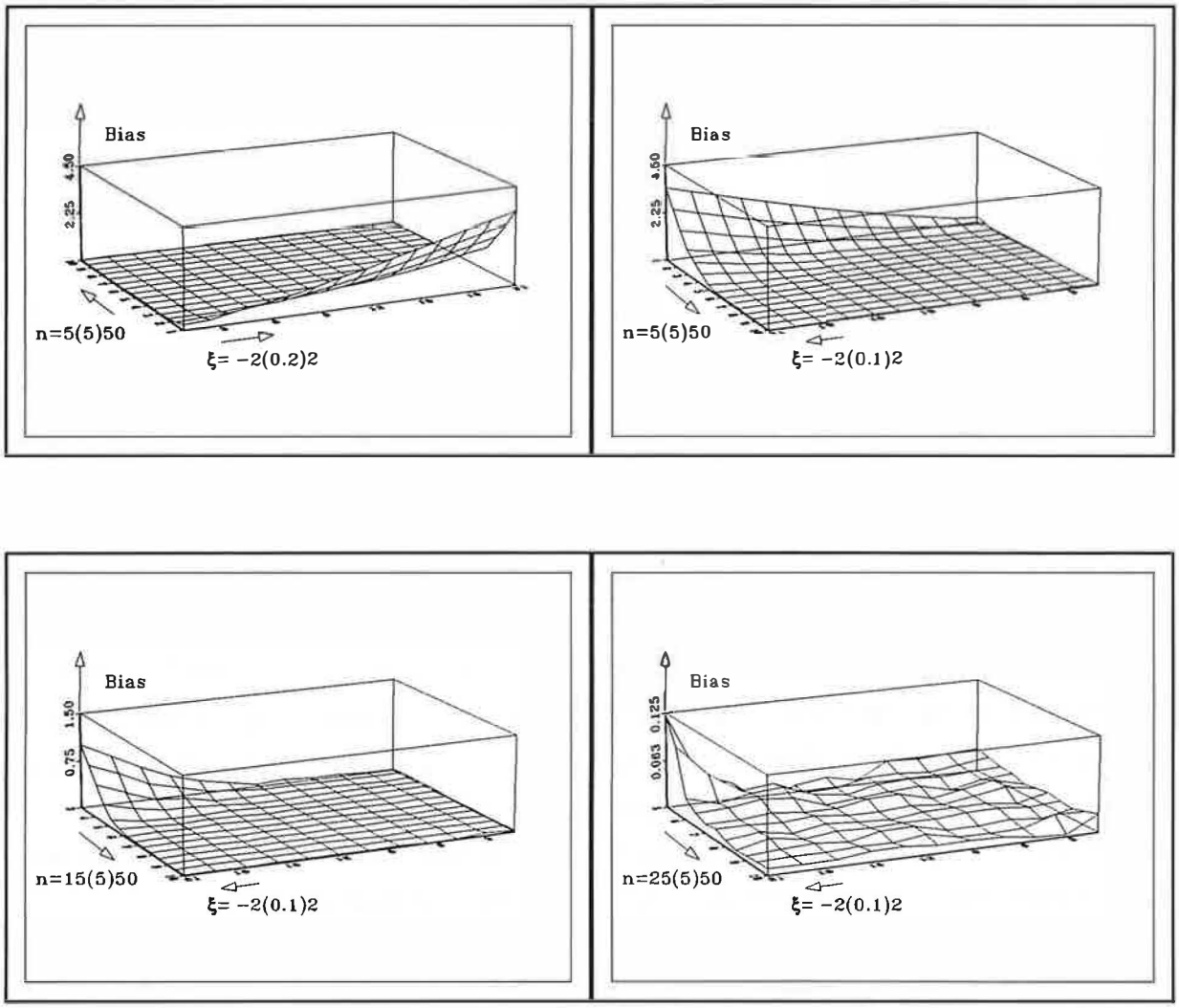

Figure 2: Bias is appreciably larger when the item difficulty increases with position in the test. Same plots as above. The fourth plot is added to show that about the same bias results from n = 25 in this condition as compared in the above from n =15.

Using $\hat{\beta} = (n+1)/\tau$ instead of $\hat{\beta} = n/\tau$ as above results in lowering the plots to beneath zero xy-plane about the same amount as they are now above for the representative part of the last plots of the two figures. Let, therefore,

$$B(x) = \frac{\xi}{\alpha} - \hat{\theta} - \ln\frac{n+x}{\tau\alpha}$$

denote the bias as a function of x. The above remarks amount to: $B(0) \approx -B(1)$, which yields:

$$\ln\frac{n}{\tau\alpha}\frac{n+1}{\tau\alpha} = 2\left[\frac{\xi}{\alpha} - \hat{\beta}\right]. \qquad (17)$$

Substitution of (17) in $B(x) = 0$ yields $n+x = \sqrt{n(n+1)}$, and, therefore, for reasonably large n: $x \approx 1/2$.

Conclusion

By using the result of Dubey, a new derivation of the Rasch model could be presented: The variation in the item responses by a given person is explained by variability in his 'momentary' ability (the random variable Z, used in (7)), and by the variation in time used to answer the item. It can be argued that there is always a certain time pressure, even in pure 'power' tests. The practice of using the Rasch model in power tests is, within the framework of the present model, nothing else than 'integrating out' the time variable.

Taking time investment into account can be accomplished in two ways, either by recording the time used per item, or by counting the number of finished items within a given time limit. In the present report, only the second approach has been studied. Of course, recording individual response times gives a lot more information and possibilities of testing the model than the mere counting of the number of finished items, but for large scale application the counting approach is more practical.

Much attention has been paid to the estimation of the parameters. It was shown that maximum likelihood is unfeasible, and that other methods have to be used. The nice result of Arnold and Strauss on pseudolikelihood estimation offers an elegant approach. Although Arnold and Strauss also give some results on the standard errors of the pseudolikelihood estimators, our elaboration of pseudolikelihood is more complicated than the cases they discuss, due to the presence of the nuisance parameters β and ϑ. Deeper study on the accuracy of the estimators as we derived them is certainly wanted.

As to the practical application of the model, the estimation of the item-parameters and the ϑ-parameters can be carried out by standard software, capable of handling incomplete designs. The estimation of β is utterly simple, as is the estimation of ξ. Also all the statistical goodness of fit tests available for testing the Rasch model are applicable here. Tests for the time distribution component of the model however are yet to be developed. At least two problems have to be thoroughly investigated: (1) Although Dubey's result allows for the whole family of gamma-distributions, only one special case was studied in the present report. The exponential distribution, and especially its constant hazard function, may be unrealistic, so that it becomes necessary to use a larger set of gamma distributions. Estimation under the general gamma distribution however is much more complicated than under the exponential distribution. It might be argued that the distribution of the response times should be dependent on the difficulty of the item. Although this may be true to a certain extent, it is not clear if this dependency should be modelled explicitly in a formal IRT-model, because then a different distribution is assumed for every subject-item combination, and nothing is left to aggregate over. This may cause severe estimation problems. The practical solution might consist in a restriction of the variability of the item difficulties, such that the dependency can be neglected for practical purposes.

Appendix A.

A small example of the calculation of the marginal probability of a response

Let $P(i, j \mid 2)$, $(i, j \in \{0, 1\}^2)$ be the probability of a response vector (i, j), given that only the first two are finished.

Thus, $P(0, 1 \mid 2)$ is the probability of a wrong response to item 1 and a correct response to item 2. The marginal probability $PM_1(0 \mid 2)$ of a wrong response to item 1 is:

$$PM_1(0 \mid 2) = P(0,0) + P(0,1) =$$

$$P(0,0|2) = \frac{2!}{\tau^2}\int_{\Delta_2} \exp(-\alpha t_1 \frac{\nu}{\rho_1}) \exp(-\alpha t_2 \frac{\nu}{\rho_2}) (dt)^2$$

$$P(0,1|2) = \frac{2!}{\tau^2}\int_{\Delta_2} \exp(-\alpha t_1 \frac{\nu}{\rho_1}) [1 - \exp(-\alpha t_2 \frac{\nu}{\rho_2})] (dt)^2$$

$$+ \quad \text{---}$$

$$PM_1(0|2) = \frac{2!}{\tau^2}\int_{\Delta_2} \exp(-\alpha t_1 \frac{\nu}{\rho_1}) (dt)^2$$

$$= \frac{2!}{\tau^2}\int_0^{\tau} \exp(-\alpha t \frac{\nu}{\rho_1}) \, dP_{\Delta_2}(T \leq t)$$

where $P_{\Delta_2}(T \leq t) = 1 - \left(1 - \frac{2}{\tau} t\right)^2$.

The ratio of the indicated surface in Δ_2 and Δ_2 itself:

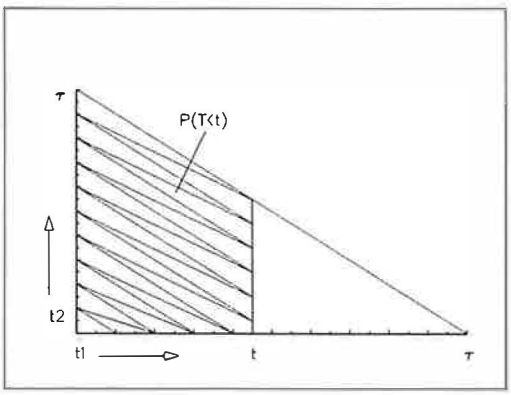

Figure 3 The probability $T \leq t$ in Δ_2

Appendix B.

The expected time per item with a uniform distribution on Δ_n of the n-dimensional vector t.

$$E_{\Delta_n}(t) = \int_0^{\tau} t dP_{\Delta_n}(T \leq t)$$

$$= \frac{n}{\tau}\int_0^{\tau} t\left[\frac{\tau - t}{\tau}\right]^{n-1} dt$$

$$= \frac{n}{\tau}\int_0^{\tau} (\tau - t)\left[\frac{t}{\tau}\right]^{n-1} dt$$

$$= \frac{\tau}{n+1}$$

Using $m = (n+1)/\tau$ in stead of $m = n/\tau = \hat{\beta}$ achieves a closer overall approximation, as shown in Figure 4 below.

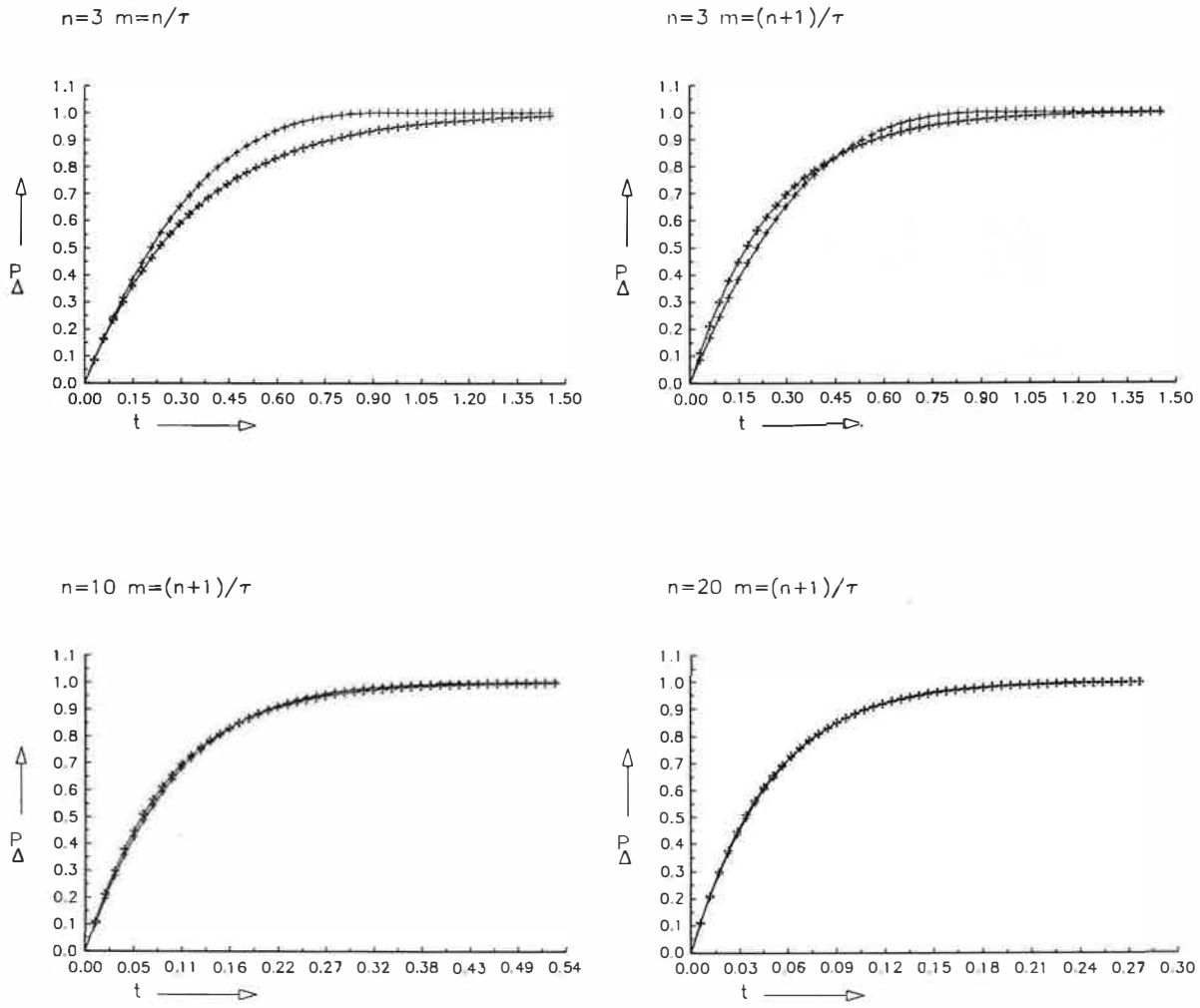

Figure 4: P_Δ and its approximations. The first two plots are for n = 3, the next for n=10 and 20. The first plot uses $m=n/\tau$ in the approximation the other plots $m=(n+1)/\tau$. The + represents P_Δ.

From simulations, however, we have the impression that $m \approx (n + \frac{1}{2})/\tau$ yields approximately unbiased estimates of ξ (see the section on bias).

Appendix C.

Sampling a time-vector $\underline{T}$ from $\Delta_{n,\tau}$.

Denote with U[a,b] the uniform distribution on the real interval [a,b]. Consider the order statistics of n independently identically $U[0,\tau]$ distributed random variables $T_i^{(n)}$, $(i=1,\ldots,n)$, and $T_{i-1}^{(n)} \leq T_i^{(n)} \leq T_{i+1}^{(n)}$, and let $T_0^{(n)} = 0$, and $T_{n+1}^{(n)} = \tau$. Then the conditional distribution of $T_i^{(n)}$, given $T_{i+1}^{(n)} = t_{i+1}^{(n)}$ is:

$$P(T_i^{(n)} \leq s \mid t_{i+1}^{(n)}) = \left[\frac{s}{t_{i+1}^{(n)}}\right]^i, \tag{18}$$

for $s \in [0, t_{i+1}^{(n)}]$, zero for smaller s and 1 for larger s. Now consider an n-dimensional $\underline{T}$ uniformly distributed on$\Delta_{n,\tau}$. More specifically, the conditional distribution of $S_i = \Sigma_1^i T_j$ given $S_{i+1} = s_{i+1}$ is (with comparable remarks for s outside the relevant interval as stated after (18)):

$$P(S_i \leq s \mid s_{i+1}) = \left[\frac{s}{s_{i+1}}\right]^i, \tag{19}$$

because of the uniform distribution of $\underline{T}$ on Δ_n, and by noting that, conditional on its sum $S_n = s$, the vector T_n is uniformly distributed on the subset of the n-1 dimensional hyperplane $\mathbb{R}^n \cap \Delta_{n,\tau}$ with $\Sigma T_i = s$. Which, by projection, is identical to a uniform T_{n-1} on $\Delta_{n-1,s}$. The same reasoning applies subsequently to $S_{n-1} = s'$ and T_{n-2} etc. down to T_1.
Formulas (18) and (19) reveal that the order statistics and the partial sums of $\underline{T}$ share the same sequence of conditional distributions, starting with the same unconditional distributions for $T_n^{(n)}$ and $\Sigma_1^n T_i$. Therefore, by taking

$$T_1 = T_1^{(n)},\ T_2 = T_2^{(n)} - T_1^{(n)},\ \ldots,\ T_n = T_n^{(n)} - T_{n-1}^{(n)},$$

one obtains an n-dimensional $\underline{T}$ uniformly distributed on Δ.

References

Amemiya, T. (1986). *Advanced econometrics*. Oxford: Blackwell.

Arnold, B.C. and Strauss, D. (1988). *Pseudolikelihood estimation.* Technical Report No. 164, Department of Statistics, University of California.

Cronbach, L.J., Warrington, W.G. (1951). Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika, 16,* 167-188.

Donders, R. (1990). *Estimation problems in the Rasch Weibull model.* (unpublished research note)

Dubey, S.D. (1969). A new derivation of the logistic distribution. *Naval Research Logistics Quarterly, 16,* 37-40.

Gullikson, H. (1950). The reliability of speeded tests. *Psychometrika, 15,* 259-269.

Guttman, L. (1955). Reliability formulas for noncompleted or speeded tests. *Psychometrika, 20,* 113-124.

Hocking, R.R. (1985). *The analysis of linear models*. Montery, Cal.: Brooks/Cole.

Iseler, A. (1970). *Leistungsgeschwindigkeit und Leistungsguete.* Weinheim: Beltz.

Johnson, L.J. and Kotz, S. (1970). *Continuous univariate distributions - 1.* New York: Wiley.

Lord, F.M. (1983). Unbiased estimation of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48,* 233-245.

Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. (Chapter 6). Reading, Massachusetts, Addison Wesley.

Meredith, William (1971). Poisson distributions of error in mental test theory. *British Journal of Mathematical & Statistical Psychology, 24,* 49-82.

Moore, William E. (1970). *Stochastic processes as true-score models for highly speeded tests. RB-70-66*, Educational Testing Service, Berkeley, Calif.

Pieters, J.P. & Van der Ven, A.H.G.S. (1982). Precision, speed and distraction time in time-limit tests. *Applied Psychological Measurement, 6,* 93-109.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press (original work published in 1960).

Roskam, E.E. (1987). Towards a psychometric theory of intelligence. In: E.E.Ch.I. Roskam & R. Suck (eds.). *Progress in mathematical psychology (vol. 1)*, p. 151-174. Amsterdam: North-Holland (Elsevier Science Publishers).

Spivak, M. (1967). Calculus. New York: Benjamin.

Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review, 34,* 273-286.

Van Breukelen G.J.P. (1989). *Concentration, speed and precision in mental tests: a psychometric approach*. Dissertation, Katholieke Universiteit Nijmegen.

Van der Ven, A.H.G.S. (1969). *The binomial error model applied to time-limit tests.* Nijmegen: Schippers.

Verhelst, N.D. and Kamphuis (1989). *Statistiek met* $\hat{\vartheta}$. Specialistisch Bulletin Nr 77. Arnhem: Cito.

Warm, Th,A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

White, P.O. (1973). Individual differences in speed, accuracy and persistence: a mathematical model for problem solving. In: Eysenck, H.J. (ed.) *The measurement of intelligence.* Lancaster, Medical and Technical Publishing Co Ltd.

Recent Measurement and Research Department Reports:

91-1 N.D. Verhelst & N.H. Veldhuijzen. A New Algorithm for Computing Elementary Symmetric Functions and Their First and Second Derivatives.

91-2 C.A.W. Glas. Testing Rasch Models for Polytomous Items: With an Example Concerning Detection of Item Bias.

91-3 C.A.W. Glas & N.D. Verhelst. Using the Rasch Model for Dichotomous Data for Analyzing Polytomous Responses.

91-4 N.D. Verhelst & C.A.W. Glas. A Dynamic Generalization of the Rasch Model.

91-5 N.D. Verhelst & H.H.F.M. Verstralen. The Partial Credit Model with Non-Sequential Solution Strategies.

91-6 H.H.F.M. Verstralen & N.D. Verhelst. The Sample Strategy of a Test Information Function in Computerized Test Design.

91-7 H.H.F.M. Verstralen & N.D. Verhelst. Decision Accuracy in IRT Models.

91-8 P.F. Sanders & T.J.H.M. Eggen. The Optimum Allocation of Measurements in Balanced Random Effects Models.

91-9 P.F. Sanders. Alternative Solutions for Optimization Problems in Generalizability Theory.

91-10 N.D. Verhelst, H.H.F.M. Verstralen & T.J.H.M. Eggen. Finding Starting Values for the Item Parameters and Suitable Discrimination Indices in the One-Parameter Logistic Model.