Measurement and Research Department Reports 2009-2

# A Poisson-Gamma model for speed tests

**N.D. Verhelst**
**F.H. Kamphuis**

Measurement and Research Department Reports 2009-2

# A Poisson-Gamma model for speed tests

N.D. Verhelst

F.H. Kamphuis

Cito
Arnhem, 2009

# A Poisson-Gamma model for speed tests

N.D. Verhelst and F.H. Kamphuis*
National Institute for Educational Measurement (Cito)
Arnhem, The Netherlands

**Abstract**

The present report discusses two applications of the Poisson measurement model for counts as originally developed by Rasch. To account for the distribution of the latent variable, a gamma distribution is hypothesized. Parameter estimation for the measurement model and for the distribution model are discussed in detail. In the first application - a collective test for technical reading - the model fits the observed data very well. In the second application - an individual test for technical reading- poor fit was found. Extending the distribution model to a mixture of two gamma variables leads to an excellent fit. Special attention is given to the estimation of the reliability.

Key words: Poisson model, gamma distribution, latent class models, reliability

## 1 Introduction

The main purpose of this report is to present an investigation on a psychometric model for pure speed tests. Speed tests can be categorized in two main types. In the first type, the testee is performing a given task, and the basic observation is the time needed to finish the task. The second type consists of performing an, in principle infinite, series of subtasks, and the basic outcome is the number of subtasks finished within a given time period. The

*We are indebted to Niels Veldhuijzen for his careful reading of the manuscript and the 106 suggestions for improving it.

subtasks themselves are easy to perform, and generally differences in difficulty between subtasks are nonexistent or ignored. This report is on tests of the latter type.

Such an investigation will comprise several parts, each of which will be given attention to. These parts are:

1. Giving some theoretical justification for the chosen model and an investigation of its mathematical features;

2. A detailed treatment of the parameter estimation procedure, with sufficient attention to cases where practical considerations necessitate the use of incomplete designs;

3. Computational aspects of the parameter estimation procedure;

4. Giving attention to the validity of the model: can it be shown that the model describes the observations to a high degree of accuracy?

5. Giving attention to the practical use of the model and its outcomes. In particular, attention must be given to the relation between a highly specialized model and a far less structured theory of measurement, Classical Test Theory, whose use is paramount among practitioners of educational testing.

To make the presentation not too abstract, and to demonstrate that the model is useful in practice, it will be applied to two complex data sets. In Section 2, the data sets will be introduced. In Section 3, the model is introduced. In Section 4, the parameter estimation problem of the Poisson part is discussed, while in Section 5, the gamma part is introduced and parameter estimation is discussed for that part. In Section 6 the relation between the present model and Classical Test Theory is discussed, with special attention to the problem of reliability. In Section 7 the results of the application of the model to one of the tests are reported. For the other test, however, the model fails to explain important features of the data. Therefore, an extension of the model is proposed. This is the subject matter of Section 8. The report concludes with a discussion (Section 9).

# 2 The tests and the data

In the student monitoring system developed by Cito for use in primary education, two different speed tests of the second type, aimed at measuring technical reading ability are used. The first one is a series of tests for collective administration; the tests are labeled 'tempo tests' (TT). The other one is administered individually and is called the 'three minutes test '(TMT). Both tests and the data collection design for the calibration are discussed next.

## 2.1 The tempo tests

The student monitoring system provides tests to primary schools in different domains, such that performances of the same student at different ages can be meaningfully compared to each other, as well as to the performances of students in the same grade. For most domains tests were provided which could be used in six of the eight grades of Dutch primary education [1]. Usually tests are administered twice a year, once in the middle of the school year and once near the end. Specific tests are indicated by an acronym designating the target grade and the period in the school year when it has to be administered. For example, a test designated for group 6 and to be administered in the middle of the school year is designated as M6, while the test intended for the end of the same grade is designated as E6. For each test norms are provided and a psychometric model is used to evaluate progress on a single underlying scale from one administration moment to the next.

For the domain of technical reading, a series of tests is provided from M4, E4, etc., to M8, and for each testing occasion a parallel test is offered as well. In the calibration study reported here a total of 19 tests has been used.

A single test consists of a coherent text where some words are left out. At each gap, three words are offered and the student has to choose the one word out of three that fits best the context. Students are instructed to underline or cross the best fitting word, and in the test instructions it is stressed that the text must be read and that students have to work as fast as they can.

[1] The Dutch system of primary education consists of eight grades (called 'groups'), the first two being equivalent to Kindergarten in many other countries. Formal instruction to reading and arithmetic starts in group 3.

Here is an example:

```
The job got out of ...  [had hand hard].
```

Each text has about one hundred gaps, and a gap appears on every line of the text. The texts are constructed to fit with the general level of reading comprehension for the grade they are intended for. This was checked by computing a readability formula (Staphorsius, 1994) on all the fully completed texts which are used in this research.

In the data collection design used to calibrate the texts and to derive the norms, each student takes two tests in a linked design. Data were collected in a period of five consecutive years in the early nineties.

As the test is meant to fit in a student monitoring system, where the reading speed can be monitored, it is unavoidable that different subtests are used at different age levels for two reasons: 1) using the same test at different ages with the same student may cause bias due to memory effects, and 2), due to the development of reading ability, texts suited for low age levels will not be usable at higher age levels and vice versa. So the model has to provide means of measuring the same concept (reading speed) using different measuring instruments.

In the model to be discussed in Section 3, it is essential that the number of words that can be read within the allotted reading time is unbounded. To prevent that students would reach the end of the text within the allotted time, the total reading time was adapted to the grade level for which the texts were constructed, ranging from 4 minutes for the higher grades to 8 minutes for the lower grades.

## 2.2 The three minutes test

The three minutes tests consists of a set of three cards, labeled 1, 2 and 3 respectively. Each card contains a list of isolated words, and the tested student is requested to read aloud each word as fast as possible but without making reading errors. The allotted time is one minute per card, and the test consists of one, two or three cards. The basic outcome is the count of the total number of words read per card and the total number of errors per card. Cards are constructed using words of similar phonological structure. The cards are referring to three types of orthographic structures: monosyllabic consonant-vowel-consonant patterns, monosyllabic words with consonant clusters and

polysyllabic words. Card 1 is intended to be the easiest and card 3 the most difficult one.

For each card three parallel forms are constructed. Parallel forms contain the same words but presented in a different order, in order to avoid rote learning.

For the calibration study, data were collected in two waves. In January 2008 data were collected in M3, M4,... up to M8 and in June 2008 data were collected again in E3 to E7, where the same students were tested as in January. At each wave each student took a form of the cards 1, 2 and 3, but the same card was never administered to the same student at the two testing occasions. Cards 1, 2 and 3 were administered in the order of increasing difficulty as this will be the practice when the test is released.

The sample sizes for the different (half) grades for the tempo tests and the three minute tests are given in Table 1.

Table 1. Sample sizes

| grade | M3 | E3 | M4 | E4 | M5 | E5 | M6 | E6 | M7 | E7 | M8 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TT | – | – | 1212 | 1513 | 810 | 857 | 854 | 864 | 837 | 792 | 655 |
| TMT | 1025 | 942 | 1018 | 920 | 954 | 811 | 879 | 765 | 779 | 704 | 775 |

# 3 The psychometric model

A very simple and parsimonious model can be derived from the following model about time investment. Suppose the time used to process a single bit of information (for the tempo tests, this means reading of the text up to the next item and responding to it, and for the TMT it just means reading the next word) is exponentially distributed with parameter $\alpha$. Denote by $S$ the number of items finished in a total time span $\tau$. Then it can be proven (Lord and Novick, 1968, pp. 490-491) that $S$ is Poisson distributed with parameter $\mu = \tau\alpha$.

Now we can let the parameter $\mu$ depend on person as well as task characteristics, i.e., we consider $\mu_{vi}$ as our basic parameters, and we decompose them as

$$\mu_{vi} = \tau_i \theta_v \sigma_i, \quad (\theta_v, \sigma_i > 0) \tag{1}$$

where $\tau_i$ is the time allotted to text $i$ or card $i$ and is expressed in an arbitrary unit; $v$ denotes the person. This model is a slight generalization of the Poisson

model developed by Rasch (1960), because it allows explicitly for variation in the time allotted to make each subtest.

Since the right hand side of (1) consists of a product of three factors, there are two opportunities to choose units of the scale. One could multiply $\tau_i$ and divide $\theta_v$ by an arbitrary positive constant $c_1$ and multiply $\theta_v$ and divide $\sigma_i$ by another arbitrary constant $c_2$ without affecting the product. One of these indeterminacies can be solved by choosing the unit of time, which in the present report will be minutes. The other determinacy is solved by choosing a normalization, which will be discussed in more detail in the next section.

This section ends with a short note on terminology. The term task will be used to refer to a text in the TT or to a card in the TMT. The term item designates a subtask in both tests, i.e. indicating a gap in the TT or a word in the TMT.

# 4 Parameter estimation

We start with the complete case: all students take the same $k$ tasks. The likelihood of the data is

$$L = \prod_v \prod_i^k \frac{\mu_{vi}^{x_{vi}}}{x_{vi}!} \exp(-\mu_{vi}) \tag{2}$$

where $v$ denotes the student, and $i$ the task; $x_{vi}$ is the value of the random variable, and indicates the number of items finished within the allotted reading time. Taking logarithms and using (1) gives

$$\begin{aligned} \ell = \ln(L) &= \sum_v \sum_i^k [x_{vi} \ln \tau_i - \ln(x_{vi}!)] \\ &\quad + \sum_v s_v \ln \theta_v + \sum_i^k t_i \ln \sigma_i \\ &\quad - \sum_v \sum_i^k \tau_i \theta_v \sigma_i \end{aligned} \tag{3}$$

where

$$s_v = \sum_i^k x_{vi} \text{ and } t_i = \sum_v x_{vi}$$

This clearly shows clearly that the model is an exponential-family model with sufficient statistics $s_v$ for the $\theta$-parameters and $t_i$ for the $\sigma$- parameters.

## 4.1 Joint Maximum Likelihood estimation (JML)

For the complete case, the estimation equations are easily derived:

$$\frac{\partial \ell}{\partial \theta_v} = \frac{s_v}{\theta_v} - \sum_i \tau_i \sigma_i, \tag{4}$$

and

$$\frac{\partial \ell}{\partial \sigma_i} = \frac{t_i}{\sigma_i} - \tau_i \sum_v \theta_v. \tag{5}$$

Of course we need a normalization (see (1)). A suitable one in the complete case is

$$\sum_i \tau_i \sigma_i = C, \tag{6}$$

with $C$ an arbitrary positive constant. So we have from (4)

$$\theta_v = \frac{s_v}{C}, \tag{7}$$

and substituting this result in (5) we find that

$$\sigma_i = \frac{C t_i}{\tau_i \sum_v s_v}. \tag{8}$$

Notice that (7) and (8) are explicit solutions; no iterations are required.

In the incomplete case, things are a little bit more involved. Defining the design indicator variables as

$$d_{vi} = \begin{cases} 1 & \text{if task } i \text{ has been administered to student } v; \\ 0 & \text{otherwise,} \end{cases}$$

the likelihood of the data is

$$L = \prod_v \prod_i^k \frac{\mu_{vi}^{d_{vi} x_{vi}}}{(d_{vi} x_{vi})!} \exp(-d_{vi} \mu_{vi}), \tag{9}$$

where $x_{vi}$ is an arbitrary number if $d_{vi} = 0$. One finds as sufficient statistics for $\theta_v$ and $\sigma_i$ respectively:

$$s_v = \sum_i d_{vi} x_{vi}, \tag{10}$$

and

$$t_i = \sum_v d_{vi} x_{vi}, \tag{11}$$

and as likelihood equations

$$\theta_v = \frac{s_v}{\sum_i d_{vi} \tau_i \sigma_i}, \tag{12}$$

and

$$\sigma_i = \frac{t_i}{\tau_i \sum_v d_{vi} \theta_v}. \tag{13}$$

This system can be solved iteratively (i.e. applying (12) and (13) alternatively), while at the same time renormalizing the $\sigma$ parameters after each evaluation of (13). Of course, the normalization is arbitrary, and a simple one, like $\prod_i \sigma_i = 1$ will do.

If the number of tasks is fixed and the number of students increases, so will the number of $\theta$ parameters, whence it is not sure that the $\sigma$ parameters are estimated consistently by using this method of estimation. Therefore we consider conditional maximum likelihood.

## 4.2 Conditional Maximum Likelihood Estimation (CML)

To investigate the conditional likelihood, it proves useful to introduce a slight reparametrizetion of the model. Define

$$\delta_i = \tau_i \sigma_i. \tag{14}$$

With a complete design, the CML estimates are easily found using a nice theorem proved by Rasch (1980) and again by Lord and Novick (1968, Theorem 21.2.4, p. 484). We state the theorem here without proof.

**Theorem 1** *Let $X_1, X_2, \ldots, X_k$ be independent Poisson random variables with parameters $\lambda_1, \lambda_2 \ldots, \lambda_k$ and let $S = X_1 + X_2 + \cdots + X_k$, then the*

*conditional distribution of $X_1, X_2, \ldots, X_k$ given $S = s$ is multinomial with index $s$ and parameters*

$$p_i = \frac{\lambda_i}{\sum_j^k \lambda_j}, \quad (i = 1, \ldots, k).$$

Use of this theorem leads immediately to the CML-estimates in a complete design. In incomplete designs, however, it is not clear at all how to use this theorem to derive the CML estimation equations. Therefore, we follow a different approach here.

Conditioning on the sufficient statistic for $\theta$, it is readily found that the conditional likelihood of the scores $(x_1, \ldots, x_k)$ on $k$ tasks is given by

$$L_c = \Pr(x_1, \ldots, x_k | s; \delta_1, \ldots, \delta_k) = \frac{\prod_i \frac{\delta_i^{x_i}}{x_i!}}{\sum_* \prod_i \frac{\delta_i^{y_i}}{y_i!}} \tag{15}$$

where $\sum_*$ runs over all $k$-tuples $(y_1, \ldots, y_k)$ of nonnegative integers such that $\sum_i y_i = s$. The denominator of (15) is a combinatorial function of $s$ and the $\delta$ parameters, which at first sight seems quite complicated, but which turns out to be utterly simple to evaluate. Define

$$\gamma_s(\boldsymbol{\delta}) \triangleq \gamma_s(\delta_1, \ldots, \delta_k) = \sum_{\sum_i y_i = s} \prod_i \frac{\delta_i^{y_i}}{y_i!} \tag{16}$$

Consider the case with $k = 2$. A total score of $s$ can arise in exactly $s + 1$ ways: $y_1$ takes the values 0 to $s$, and $y_2$ takes the values $s - y_1$. If $y_1$ takes the value $j$, the corresponding term in the sum of (16) is

$$\frac{\delta_1^j}{j!} \times \frac{\delta_1^{s-j}}{(s-j)!}$$

whence we obtain

$$\begin{aligned} \gamma_s(\delta_1, \delta_2) &= \sum_{j=0}^{s} \frac{\delta_1^j}{j!} \times \frac{\delta_2^{s-j}}{(s-j)!} \\ &= \frac{1}{s!} \sum_{j=0}^{s} \binom{s}{j} \delta_1^j \delta_2^{s-j} \\ &= \frac{1}{s!} (\delta_1 + \delta_2)^s \end{aligned} \tag{17}$$

To generalize to the case with arbitrary $k$, we need the following theorem.

**Theorem 2** $\gamma_s(\delta_1, \ldots, \delta_k) = \frac{1}{s!}\left[\sum_i^k \delta_i\right]^s$

**Proof.** The proof is by induction. It trivially holds for $k = 1$. The induction hypothesis is that it holds for $k - 1$. The variable $y_k$ in the sum of (16) can take the values 0 through $s$. If it takes the value $j$, the sum of the other $y$ values is $s - j$, and these values can be distributed in an number of ways over the values $y_1$ through $y_{k-1}$. But the contribution of all these possibilities is exactly $\gamma_{s-j}(k-1)$. Therefore, if $y_k = j$, the contribution to $\gamma_s(k)$ is

$$\frac{\delta_k^j}{j!} \times \gamma_{s-j}(k-1),$$

which is, by the induction hypothesis, equal to

$$\frac{\delta_k^j}{j!} \times \frac{\left[\sum_i^{k-1} \delta_i\right]^{s-j}}{(s-j)!}.$$

Therefore

$$\begin{aligned}
\gamma_s(\delta_1, \ldots, \delta_k) &= \sum_{j=0}^{s} \frac{\delta_k^j}{j!} \times \frac{\left[\sum_i^{k-1} \delta_i\right]^{s-j}}{(s-j)!} \\
&= \frac{1}{s!}\left[\delta_k + \sum_i^{k-1} \delta_i\right]^s \\
&= \frac{1}{s!}\left[\sum_i^{k} \delta_i\right]^s. \qquad (18)
\end{aligned}$$

■

It is easy to see that $\gamma_0(\boldsymbol{\delta}) = 1$. To have consistent notation, we define

$$\gamma_s(\boldsymbol{\delta}) = 0 \text{ if } s < 0. \qquad (19)$$

Using (18), one obtains a useful recursive relation:

$$\gamma_s(\boldsymbol{\delta}) = \gamma_{s-1}(\boldsymbol{\delta}) \times \frac{\sum_i^k \delta_i}{s} \qquad (20)$$

from which it follows immediately that

$$\gamma_s(\boldsymbol{\delta}) \lesseqgtr \gamma_{s-1}(\boldsymbol{\delta}) \Longleftrightarrow \textstyle\sum_i^k \delta_i \lesseqgtr s \tag{21}$$

and it follows that $\gamma_s(\boldsymbol{\delta})$, considered as a function of $s$, is either monotonously decreasing (if $\sum_i^k \delta_i < 1$), takes the value 1 for $s = 0$ and $s = 1$ and then decreases (if $\sum_i^k \delta_i = 1$), or is single-peaked, since $s$ is unbounded and $\sum_i^k \delta_i$ is independent of $s$.

For computational purposes, it may be useful to have a rough estimate of the order of magnitude of the combinatorial function $\gamma$. From (18) and the Taylor expansion of the exponential function, it follows immediately that

$$\sum_{s=0}^{\infty} \gamma_s(\boldsymbol{\delta}) = \exp\left[\textstyle\sum_i^k \delta_i\right] \tag{22}$$

This relation may be useful in choosing a suitable normalization.

From (18) it follows that

$$\frac{\partial}{\partial \delta_i}\gamma_s(\boldsymbol{\delta}) = \gamma_{s-1}(\boldsymbol{\delta}). \tag{23}$$

To derive the likelihood equations in case of an incomplete design, define the vector $\boldsymbol{\delta}_v$ as the vector of $\delta$-parameters belonging to the tasks which have been administered to student $v$. From (15), we immediately have that the conditional log-likelihood for a single subject $v$ is

$$\ell_{cv} = \sum_i^k x_{vi} d_{vi} \ln \delta_i - \ln \gamma_{s_v}(\boldsymbol{\delta}_v) \tag{24}$$

where $s_v$ is the score of student $v$, defined by (10). Now assume that student $v$ has answered to task $i$, where $i$ denotes a specific task, then we find

$$\frac{\partial \ell_{cv}}{\partial \delta_i} = \frac{x_{vi}}{\delta_i} - \frac{\gamma_{s_v-1}(\boldsymbol{\delta}_v)}{\gamma_{s_v}(\boldsymbol{\delta}_v)}$$

and for all students who have responded to task $i$, we find, using (11) and (23) that

$$\begin{aligned} \frac{\partial \ell_c}{\partial \delta_i} &= \frac{\sum_v x_{vi} d_{vi}}{\delta_i} - \sum_v d_{vi} \frac{\gamma_{s_v-1}(\boldsymbol{\delta}_v)}{\gamma_{s_v}(\boldsymbol{\delta}_v)} \\ &= \frac{t_i}{\delta_i} - \sum_v d_{vi} s_v \left[\mathbf{1}'\boldsymbol{\delta}_v\right]^{-1} \end{aligned} \tag{25}$$

From (25) the likelihood equations immediately follow:

$$\delta_i = \frac{t_i}{\sum_v d_{vi} s_v \left[\mathbf{1}'\boldsymbol{\delta}_v\right]^{-1}}, \qquad (i = 1, \ldots, k) \tag{26}$$

## 4.3 The relation between CML and JML

Taking into account the reparameterization (14), the JML equation for $\delta_i$ can be written as

$$\delta_i = \frac{t_i}{\sum_v d_{vi}\theta_v}. \tag{27}$$

Substituting (12) for $\theta_v$ one obtains

$$\delta_i = \frac{t_i}{\sum_v d_{vi} s_v \left[\sum_j d_{vj}\delta_j\right]^{-1}} \tag{28}$$

but this is the same as (26), whence it follows that JML and CML yield the same estimates for $\delta_i$ (and hence for $\sigma_i$). So the procedures using JML or CML for the task parameters, followed by a maximum likelihood estimation of the $\theta$ parameters (with the $\sigma$ parameters fixed at their CML estimates) lead to identical results. This is quite different from the Rasch model for binary responses.

## 4.4 Estimation of $\theta$

Since the CML-estimates of the $\sigma$-parameters are identical to the JML-estimates, the estimates of $\theta$ at the CML-estimates of $\sigma$ are identical to the JML-estimates of $\theta$. Defining

$$\widehat{\delta}_v = \sum_i d_{vi}\tau_i\widehat{\sigma}_i, \tag{29}$$

and using (12), we find that

$$\widehat{\theta}_v = \frac{s_v}{\widehat{\delta}_v}. \tag{30}$$

Using (4), we find that

$$-E\left[\frac{\partial^2 \ell}{\partial \theta_v^2}\right] = \frac{E(s_v)}{\theta_v^2} = \frac{\delta_v}{\theta_v},$$

from which we find that

$$SE(\widehat{\theta}_v) = \sqrt{\frac{\theta_v}{\delta_v}} \approx \sqrt{\frac{\widehat{\theta}_v}{\widehat{\delta}_v}} = \frac{\sqrt{s_v}}{\widehat{\delta}_v}. \tag{31}$$

If $\widehat{\delta}_v = \delta_v$, then the estimator (30) is conditionally unbiased as can easily be seen:

$$E(\widehat{\theta}|\theta) = \frac{1}{\delta} E(S|\theta) = \frac{1}{\delta} \times \delta\theta = \theta \tag{32}$$

# 5 The population model

The estimation problem discussed in the previous section are of limited value when it comes to construct norm tables. The term 'norm tables' is in fact synonymous with the distribution of the measured variable (displayed in a table). Tables of the 99 percentiles P1 to P99 are very common. But an important question is which variable will be tabulated. In the present context one might wish to determine the distribution of the reading ability, $\theta$, or the distribution of the estimated reading ability $\widehat{\theta}$. In the latter approach, two ways can be followed: either using JML and obtaining task and person parameters at the same time, or using CML to estimate the task parameters, then fixing these parameters at their estimates and obtaining ML estimates of the person parameters. Both procedures lead to the same estimates for task and person parameters.

Basing norm tables on the estimated person parameters from a representative sample from some population will give a consistent estimate of the percentiles of the estimated theta values **for the test forms which have been used in the calibration research**. But if in the application of the test in real life, other test forms are used, these percentiles will become biassed. Here is a concrete example: in the calibration research of the TT, each student has taken two texts (in order to realize a connected design), but in the practical application of the test, the administration of the TT will only use a single text, with the consequence that the estimated theta values of real applications will have a larger standard error than the ones based on two texts. Using the general formula of variance decomposition,

$$Var(\widehat{\theta}) = Var[E(\widehat{\theta}|\theta)] + E[Var(\widehat{\theta}|\theta)] \tag{33}$$

and using (32), the first term in the right-hand side of (33) is the variance of the true person parameters, and the second term is the average error variance,

which will vary depending on which and how many tests are administered. Therefore we prefer to estimate the distribution of the latent variable $\theta$ itself.

A well known method to estimate the percentiles is to consider the person parameters $\theta$ as realizations of a random variable. The assumed distributional form of this random variable is the population model, and in an empirical context this distribution must be estimated. Two distribution families are frequently used to model positive continuous variables: the log-normal and the gamma distribution. As the gamma distribution and the Poisson distribution go well together (the gamma is the conjugate of the Poisson distribution), we will pay attention to the gamma distribution in the present section.

## 5.1 The gamma distribution

The probability density function (pdf) of the gamma distribution is given by

$$g(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta), \quad (\alpha, \beta > 0). \tag{34}$$

where $\theta$ is the random variable and $\alpha$ and $\beta$ are the parameters of the distribution. $\Gamma(.)$ is the gamma function, and can be considered as an extension of the factorial function to continuous arguments. When the argument is a positive integer, it holds that

$$\Gamma(\alpha) = (\alpha - 1)!$$

The relation between the two parameters and the first moments of the gamma distribution are simple:

$$E(\theta) = \frac{\alpha}{\beta}, \tag{35}$$

and

$$Var(\theta) = \frac{\alpha}{\beta^2}. \tag{36}$$

## 5.2 The marginal distribution of the scores

Suppose the random variable $S$ is the total number of tasks finished by some student when administering him or her $k$ texts in a total time of $\sum_{i=1}^{k} \tau_i$ units.

The measurement model (the Poisson model) assumes that the distribution of $S$ is Poisson with parameter

$$\theta \times \sum_{i}^{k} \tau_i \sigma_i,$$

where the quantities $\tau$ are assumed to be known, and the difficulty parameters $\sigma$ can be estimated by CML. As is done in applications of latent regression, the parameters $\sigma$ will be treated as known and fixed at their CML-estimates. We will use the shorthand notation

$$\delta = \sum_{i}^{k} \tau_i \sigma_i. \tag{37}$$

The marginal probability of $s$ is then given by

$$f(s) = \int_0^\infty \frac{(\delta\theta)^s \exp(-\delta\theta)}{s!} \times \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta)\ d\theta. \tag{38}$$

Multiplying and dividing the right-hand side of (38) by

$$(\delta + \beta)^{s+\alpha} \times \Gamma(\alpha + s)$$

makes it possible to get rid of the integral, giving as a result

$$f(s) = \frac{\Gamma(\alpha + s)}{s!\Gamma(\alpha)} \times \frac{\delta^s \beta^\alpha}{(\delta + \beta)^{s+\alpha}}. \tag{39}$$

Defining

$$p = \frac{\delta}{\delta + \beta},$$

equation (39) can be written as

$$f(s) = \frac{\Gamma(\alpha + s)}{s!\Gamma(\alpha)} p^s (1 - p)^\alpha, \tag{40}$$

which is the negative binomial distribution, also known as the Gamma-Poisson distribution.

Although it is possible to start from (40) to find the MML estimates of $\alpha$ and $\beta$, and the text parameters $\delta$ at the same time, the estimation procedure in the case of incomplete designs (with texts of different difficulty)

and different reading times, leading to different values of the parameter $\delta$ in different cells of the design, is quite involved. Instead, we will fix the $\delta$ parameters at their CML-estimates, and use the marginal distribution only for estimating the $\alpha$ and $\beta$-parameters. This implies that the estimates of the population parameters only depend on the total score that a student has obtained on all tasks which have been administered to him of her. Moreover, we will not use the form (40) but (39) and simplify it further to get rid of the $\Gamma$-functions.

Using the recurrence relation

$$\Gamma(z+1) = z\Gamma(z)$$

and taking into account that $s$ is a non-negative integer, we find as an explicit expression without use of the $\Gamma$-functions:

$$f(s) = \frac{\delta^s \beta^\alpha}{s!(\delta+\beta)^{s+\alpha}} \times \prod_{i=0}^{s-1}(\alpha+i). \tag{41}$$

The product in the right-hand side of (41) equals 1 if $s = 0$.

To compute the distribution, we notice from (39) that

$$f(0) = \left[\frac{\beta}{\delta+\beta}\right]^\alpha = (1-p)^\alpha$$

and that

$$f(s) = f(s-1) \times p \times \frac{\alpha+s-1}{s}, \quad (s>0). \tag{42}$$

Using (42) in the infinite sum $\sum_{x=0}^{\infty} sf(s)$, and using the fact that $\sum_{x=0}^{\infty} f(s) = 1$, we readily find the relation

$$E(S) = p\alpha + pE(S),$$

whence it follows that

$$E(S) = \alpha\frac{p}{1-p} = \frac{\alpha\delta}{\beta}. \tag{43}$$

For the variance, the result is

$$\mathrm{Var}(S) = \alpha\frac{p}{(1-p)^2} = \frac{\alpha\delta}{\beta}\left(1+\frac{\delta}{\beta}\right) \tag{44}$$

These two moments can be used to find suitable moment estimators of $\alpha$ and $\beta$ as starting values for the MML procedure. Moreover, notice that the variance is larger than the mean, showing clearly that the negative binomial is a good candidate for explaining overdispersion phenomena in Poisson models.

## 5.3 Estimation procedure

The log-likelihood function is just the logarithm of (41) and equals

$$\ln L(\alpha, \beta; s) = C + \alpha \ln(\beta) + \sum_{i=0}^{s-1} \ln(\alpha + i) - (s + \alpha) \ln(\delta + \beta). \tag{45}$$

where $C = -\ln(s!)$ does not depend on the parameters. The partial derivatives are

$$\frac{\partial \ln L}{\partial \alpha} = \ln \frac{\beta}{\delta + \beta} + \sum_{i=0}^{s-1} \frac{1}{\alpha + i},$$

(where the sum equals zero if $s = 0$), and

$$\frac{\partial \ln L}{\partial \beta} = \frac{\alpha}{\beta} - (s + \alpha) \frac{1}{\delta + \beta}.$$

For the second partial derivatives we find

$$\frac{\partial^2 \ln L}{\partial \alpha^2} = -\sum_{i=0}^{s-1} \frac{1}{(\alpha + i)^2},$$

$$\frac{\partial^2 \ln L}{(\partial \beta)^2} = -\frac{\alpha}{\beta^2} + (s + \alpha) \frac{1}{(\delta + \beta)^2}$$

and

$$\frac{\partial^2 \ln L}{\partial \alpha \partial \beta} = \frac{1}{\beta} - \frac{1}{\delta + \beta}$$

Using the moment estimators as starting values and applying the Newton-Raphson algorithm to solve the likelihood equations did not give any problems in all the applications. Usually two or three iterations were sufficient to give estimates accurate to about ten decimal digits.

## 5.4 Estimation of $\theta$ (revisited)

The marginal model gives an opportunity to estimate $\theta$ as the expected a posteriori ability level. The posterior distribution of $\theta$ given the score $s$ is also gamma:

$$\theta | s \sim \text{Gamma}(\alpha + s, \beta + \delta)$$

from which it is immediately found that

$$E(\theta | s) = \frac{\alpha + s}{\beta + \delta}. \tag{46}$$

The expected a posteriori (EAP) value of $\theta$ can be used as an alternative estimator of $\theta$, instead of (30), the JML-estimator. The relation between the EAP-estimator and the JML-estimator $\widehat{\theta}$ is easily checked: it holds that

$$E(\theta|s) \lesseqgtr \widehat{\theta} \Leftrightarrow s \gtreqless \frac{\alpha\delta}{\beta} = E(S)$$

which shows that there is a shrinking towards the mean. This also means that the EAP-estimator is (conditionally) biassed.

The posterior standard deviation, which can be used as a substitute for the standard error is

$$SD(\theta|s) = \frac{\sqrt{\alpha + s}}{\beta + \delta} \tag{47}$$

The following relation holds:

$$SD(\theta|s) < SE(\widehat{\theta}) \Leftrightarrow s > \frac{\alpha\delta^2}{\beta(\beta + 2\delta)} = E(s) \times \frac{\delta}{\beta + 2\delta}$$

which shows that the posterior standard deviation is larger than the standard error of the JML-estimate only for relatively small values of the score.

# 6 The reliability of the speed tests

Given the latent value of $\theta$, the score distribution is Poisson with parameter (and expected value) $\delta\theta$. Its variance also equals $\delta\theta$, and can be considered as the variance of the measurement error. Since $\theta$ is assumed to be gamma distributed (with parameters $\alpha$ and $\beta$), the expected error variance is

$$E[Var(S|\theta)] = \delta E(\theta) = \frac{\alpha\delta}{\beta} \tag{48}$$

Combining this with (44), we find

$$\rho_{SS\prime} = 1 - \frac{E[Var(S|\theta)]}{Var(S)} = 1 - \frac{1}{1 + \frac{\delta}{\beta}} = \frac{\delta}{\delta + \beta} = p \tag{49}$$

The reliability of the JML-estimates $\widehat{\theta}$ is the same, because $\widehat{\theta}$ is proportional to the score $S$ (see equation 30.) The expression $\delta/(\delta+\beta)$ clearly shows

that the reliability of a test depends on characteristics of the test (through the parameter $\delta$) as well as on characteristics of the population where the test is to be applied (through the parameter $\beta$). If the population is fixed, the only way to influence the reliability is to alter the test such that its $\delta$-parameter changes. It is well known that the standard way of increasing the reliability of a test is making it longer. In the case of the speed tests discussed here, this means increasing the allotted reading time, and at the same time lengthening the task (text or card) such that the end is not reached in the lengthened reading time. If the reading time is increased by a factor $f$, then the reliability becomes

$$\rho_{SS\prime}(f) = \frac{f\delta}{f\delta + \beta} = \frac{f\rho_{SS\prime}}{1 + (f-1)\rho_{SS\prime}}$$

which is the well-known Spearman-Brown formula.

Good use of the formula (49) can be made to estimating the correlation between two latent concepts where the same kind of model is used, by applying the correction for attenuation on the observed correlation between the two series of scores. This avoids using a multivariate gamma distribution.

# 7 Results for the tempo tests

This section is divided into three subsections. In the first one, results are presented for the $\sigma$-parameters. Next the results for the population models are presented and in the final subsection the problem of the validity of the model is addressed. There it will appear that the theoretical model discussed thus far, is valid for the tempo tests and can be trusted for applications.

## 7.1 The difficulty of the tempo test forms

In Figure 1 the CML-estimates of the $\sigma$-parameters are displayed graphically. The labels along the horizontal axis indicate the grade for which the texts were originally constructed. The estimates are displayed along the vertical

axis. The product of the 19 estimates equals one.

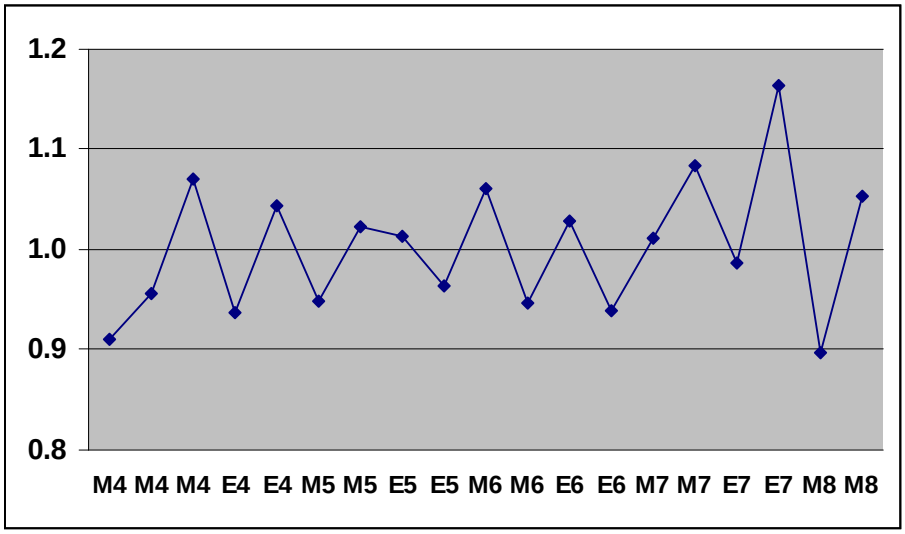


Figure 1. $\sigma$-estimates for the 19 texts of the tempo tests

At first sight it may be puzzling that there is no clear indication of a trend in the estimates. The two most difficult texts - the ones with the lowest $\sigma$-value - are constructed for M4 (the first) and for M8 (the last but one) respectively. To understand the power and at the same time the limits of the model used, two aspects must be taken in consideration. The first is a clear interpretation of the parameters of the model; the second one concerns the data collection design.

The basic parameter of the psychometric model, $\mu_{vi}$, is the mean of the Poisson distribution, and its dimension is therefore the same as the basic observation which is a frequency, and might be labeled as 'number of subtasks completed'. In the model the Poisson parameter is decomposed as a product of three factors:

$$\mu_{vi} = \tau_i \sigma_i \theta_v$$

and it may be useful to assign a dimension to each of the three factors. Here is how one can do this:

1. $\tau_i$ is a time, and the unit is free, but we have chosen minutes as the unit;

2. $\sigma_i$ is a dimensionless number, called 'impediment' by Rasch (1980, pp. 17). A text with a $\sigma$-value of 1 could be referred to as a standard

text. It the sigma value of a particular text is less than one, then its impediment is greater than the one of a standard text and this will lead to less subtasks completed;

3. Since $\mu_{vi}$ is the number of subtasks completed, it must follow that $\theta_v$ is the number of subtasks completed per unit of time, i.e., a measure of reading speed.

Now suppose two students, one of grade 4 and one of grade 8, have the same reading speed. The model then implies that their expected performances on whatever text is the same, but given the construction principles of the texts, this is a highly unrealistic assumption. The texts constructed for grade 8 usually will contain words and grammatical structures which are in general inaccessible to grade 4 students, and conversely, it may seem quite unrealistic to assume that a grade 8 student - with five to six years of formal instruction in reading - goes through a very simple text in the same way as a young student with quite limited experience in reading.

In principle, these implications can be tested empirically, but such a test would imply bringing students in a very unnatural situation, causing all kinds of special effects (such as frustration, boredom, a feeling of humiliation, etc...) which in all probability would interfere with the concept of reading speed. In the data collection design, therefore, texts were only administered to students of the same grade or a neighboring grade which they were constructed for. The extent to which the kind of extrapolations (such as what will a grade 4 student do on a text aimed at group 8) discussed in the previous paragraph hold or do not hold therefore remain unanswered.

## 7.2 The population parameters

For the nine grade-and-period combinations (M4 to M8), a gamma distribution has been estimated, using the CML-estimates of the $\sigma$-parameters as fixed constants. In Table 2, the estimates together with their estimated standard errors are displayed. In the last two columns estimated average and standard deviation, using (35) and (36) are displayed as well.

Table 2. Parameter estimates of the tempo test distributions

| grade | $\alpha$ | $SE(\alpha)$ | $\beta$ | $SE(\beta)$ | Mean | SD |
|---|---|---|---|---|---|---|
| M4 | 9.99 | 0.45 | 1.81 | 0.08 | 5.53 | 1.75 |
| E4 | 10.31 | 0.41 | 1.64 | 0.07 | 6.29 | 1.96 |
| M5 | 14.24 | 0.81 | 2.12 | 0.12 | 6.71 | 1.78 |
| E5 | 15.08 | 0.83 | 2.02 | 0.11 | 7.48 | 1.93 |
| M6 | 14.19 | 0.77 | 1.70 | 0.09 | 8.37 | 2.22 |
| E6 | 18.14 | 1.02 | 1.98 | 0.11 | 9.16 | 2.15 |
| M7 | 16.88 | 0.95 | 1.72 | 0.10 | 9.82 | 2.39 |
| E7 | 17.67 | 1.03 | 1.77 | 0.10 | 9.96 | 2.37 |
| M8 | 19.85 | 1.34 | 1.93 | 0.13 | 10.27 | 2.31 |

The cumulative distributions of the reading speed variable are displayed in Figure 2. The black curves are the distributions at the medium moment, the grey ones at the end moments; the curves are neatly ordered in the same way as the means displayed in Table 2. The general trend of a larger variation with increasing grades (see the column 'SD' in Table 2) is depicted by curves becoming flatter, the higher the grade.

The most remarkable feature, however, of Figure 2 is the very small progress in performance for the four highest groups (representing a time span of full two years) for the very weak readers (the lowest 5%, say). Percentile 5 is 5.9 subtasks per minute at M6 and only 6.8 at M8.

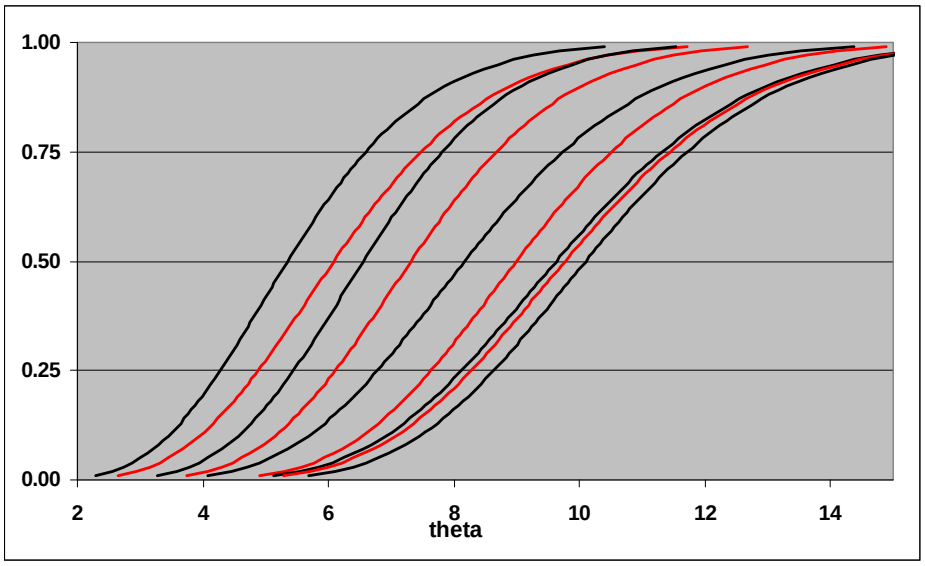


Figure 2. Cumulative distributions of the reading speed for TT (M4 to M8)

In Figure 3, five percentiles of the distributions are displayed. From bottom to top: P10, P25, the median, P75 and P90. The dashed line through the medians is a linear trend line, showing that the estimated medians tend to level off a bit in the two highest grades, although perhaps less than one would expect if the tempo tests are to be interpreted as tests of technical reading.

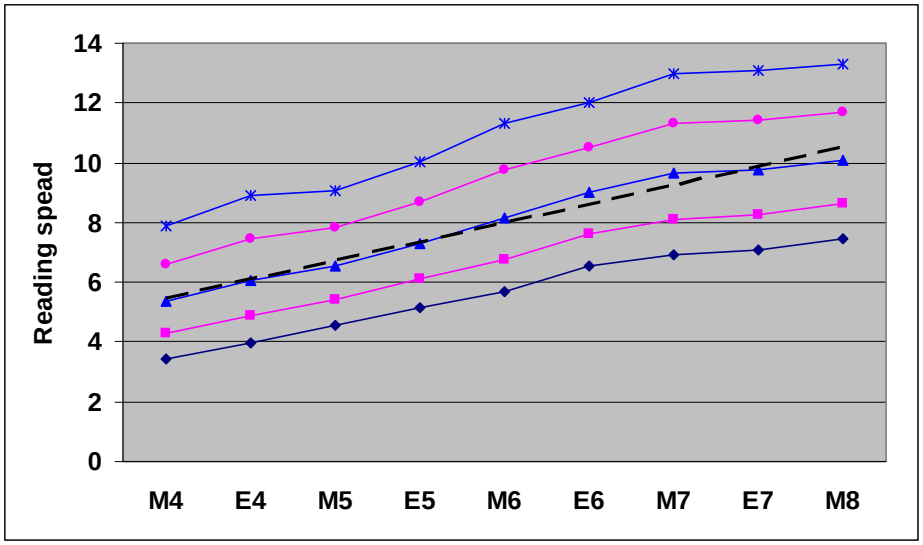


Figure 3. Percentiles 10, 25, 50, 75 and 90 for the reading speed in the tempo tests

## 7.3 The validation of the model for the tempo tests

A powerful and very elementary way of validating the model, especially in the case where percentiles of the ability distribution have to be estimated, is by looking at the accuracy of predicting the observed distribution of the scores. In Figure 4, the frequency polygons of the results for the E4 sample, observed and predicted frequencies are displayed. Since 150 score points are displayed, and the total sample size for this population is about 1500, it is to be expected that the observed frequency polygon will show many

irregularities, such that it is quite difficult to judge the fit of the model.

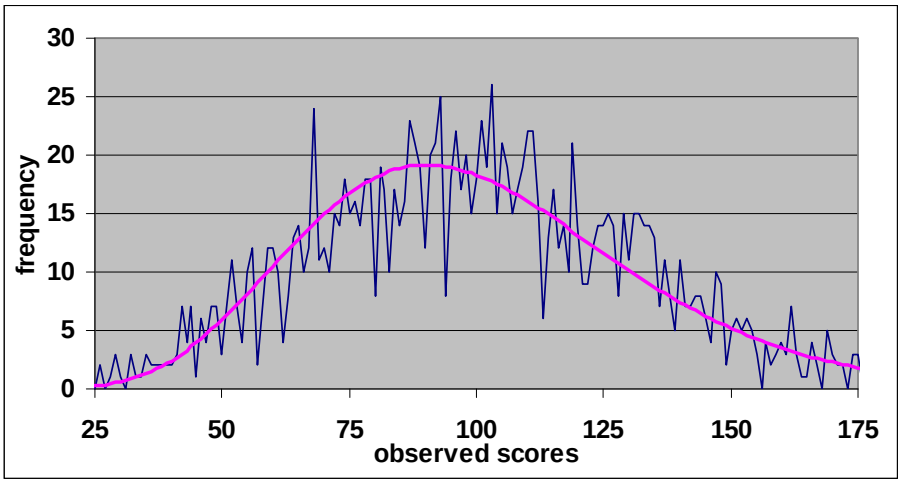


Figure 4. Observed and predicted frequencies for E4

To get rid of this irregular appearance, one can use cumulative frequency polygons. Observed and expected polygons for M4 and E4 are displayed in Figure 5. The left two curves apply to M4, and virtually coincide, while for E4 the fit is not as good. These two examples represent a best and a worst case for the nine populations (M4 to M8) which have been estimated. In general, however, the fit is satisfactory, and is a good basis to trust the estimated distributions which are displayed graphically in Figure 2.

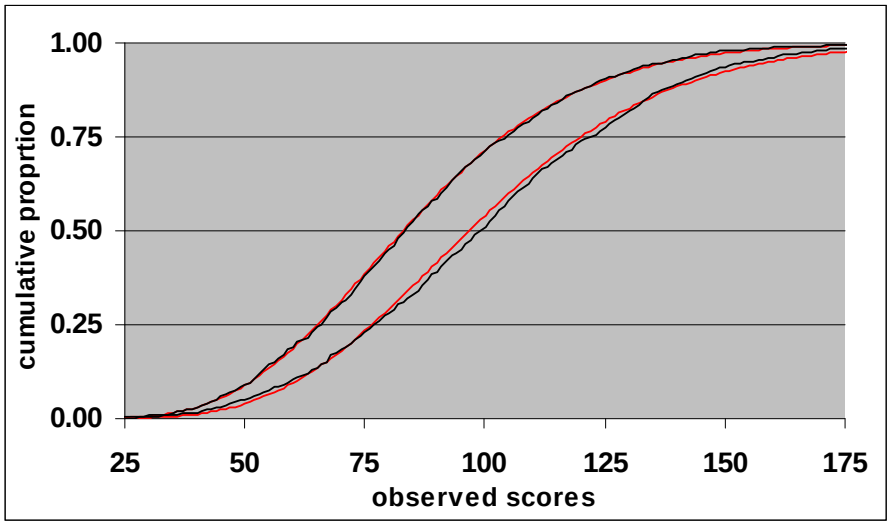


Figure 5. Cumulative distributions for M4 and E4

A good statistical fit of the model, however, does not imply an optimal quality for the use of the tests in individual cases. In Table 3, the reliabilities of the individual texts (considered as a single test) are given for the populations for which they were originally constructed: three texts for M4, and two texts for the remaining measurement moments. In the second column the reading time for the texts is displayed. Although the trend is not linear, it is clear that the reliabilities drop as the reading time decreases. For individual texts, the reliability might be deemed too low, but it can be increased by letting students take two or more texts.

Table 3. Reliabilities of the individual texts

| grade | $\tau$ | | | |
|---|---|---|---|---|
| M4 | 8 | 0.80 | 0.81 | 0.83 |
| E4 | 8 | 0.82 | 0.84 | |
| M5 | 7 | 0.76 | 0.77 | |
| E5 | 7 | 0.78 | 0.77 | |
| M6 | 6 | 0.79 | 0.77 | |
| E6 | 6 | 0.76 | 0.74 | |
| M7 | 5 | 0.75 | 0.76 | |
| E7 | 5 | 0.74 | 0.77 | |
| M8 | 4 | 0.69 | 0.65 | |

The values in Table 3 are computed using formula (49). Take the first test constructed for grade M4 as an example. Its $\sigma$- value is 0.91, so that the associated $\delta$-value is $8 \times 0.91 = 7.28$. The $\beta$-parameter estimate is 1.81 (see Table 2), and applying formula (49) yields

$$\rho_{SS'} = \frac{7.28}{7.28 + 1.81} = 0.80.$$

# 8 Results for the three minute tests

The three minutes test has been administered to a calibration sample ranging from the grades M3 to M8. As the basic outcome of this test is also a count (the number of words correctly read), the same model has been applied as with the tempo tests. The results, however, were disappointing, as may be seen from Figure 6, where the distributions of observed and predicted scores

are displayed for M3.

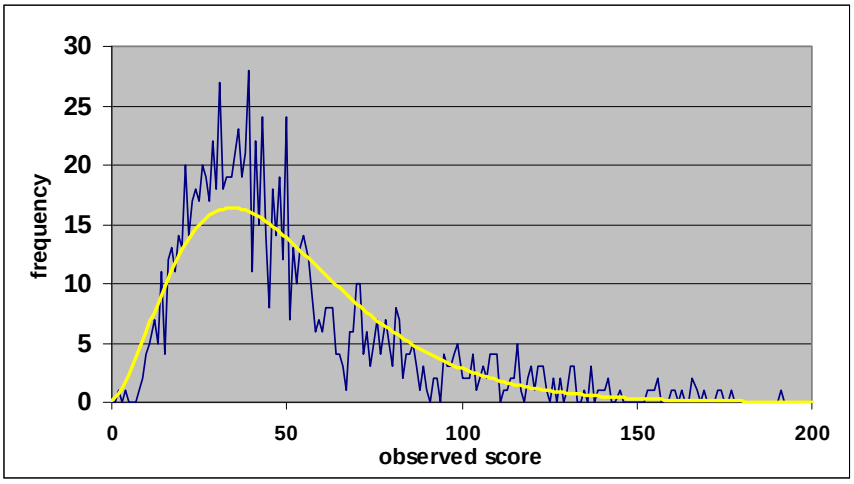


Figure 6. Validation of the Poisson-Gamma model for M3

The observed distribution shows at the same time a rather tight concentration of observations around the modal score (around 40) and a rather heavy tail for scores larger than 75, say. The theoretical model apparently is not capable of grasping these two features at the same time. A possible solution of this problem is to conceive of the M3 population as a composition of two unknown subpopulations or latent classes. This idea is elaborated in the next section.

## 8.1 A latent class model

The basic idea is that the measurement model (the Poisson model) is valid for all members of the total population. This means that each task is characterized by a single $\sigma$-parameter. The distribution of the latent variable (reading speed), however, differs for the two classes. The concrete assumption we are working with is that in both classes the latent variable is gamma-distributed, but with different parameters. As the classes are not identified, we have no ready-made estimate of the number of students in the sample belonging to each class. We will assume that a proportion $\pi_1$ of the population belongs to class 1, and the remaining proportion $\pi_2 = 1 - \pi_1$ belongs to class 2.

In summary then five parameters have to be estimated from the data: $\alpha_1$ and $\beta_1$, the gamma distribution parameters for class 1; $\alpha_2$ and $\beta_2$ for the gamma distribution in class 2, and the mixing proportion $\pi = \pi_1$. The

EM-algorithm (Dempster, Laird and Rubin, 1977) is very well suited for estimating these parameters.

## 8.2 The EM-algorithm

In this section we give a rather informal description of the EM algorithm as applied to the present problem.

If we knew for each student whether he/she belongs to class 1 or to class 2, the problem would be quite simple. The best estimate for the mixing proportion $\pi$ is the proportion (in the sample) of students belonging to class 1. For the estimation of the parameters of the two gamma distributions, one could proceed in the same way as with the model with one gamma distribution: for each subsample the gamma distribution for the corresponding class is estimated. The total computational load would then be about the double of the case with a single distribution.

If we do not know to which class each student belongs, then we have to estimate this in some way. The procedure can be described by the following scheme:

**step 0 (Initialisation)** Find some suitable values for the five parameters, and label them $\widetilde{\alpha}_1$, $\widetilde{\beta}_1$, $\widetilde{\alpha}_2$, $\widetilde{\beta}_2$ and $\widetilde{\pi}$. These values are called the current estimates of the parameters.

**step 1 (E-step)** Using $\widetilde{\alpha}_1$, $\widetilde{\beta}_1$, $\widetilde{\alpha}_2$, $\widetilde{\beta}_2$ and $\widetilde{\pi}$, compute the conditional probability for each student that he belongs to class 1 (or class 2), given his score $s$ on the test. This probability is denoted as

$$\widetilde{P}(C = c|s), \quad (c = 1, 2)$$

where we use the symbol $\widetilde{P}$ to indicate that we have to use the current value of the parameter estimates to compute these probabilities. Then the expected number of students having the score $s$ and belonging to class $c$ is given by

$$n_s \widetilde{P}(C = c|s)$$

where $n_s$ is the observed number of students with score $s$.

**step 2 (M-step)** Perform the analysis as if the expected frequencies computed in the previous step were the observed frequencies. The outcome of this analysis are new parameter estimates which we denote as $\widehat{\alpha}_1$, $\widehat{\beta}_1$, $\widehat{\alpha}_2$, $\widehat{\beta}_2$ and $\widehat{\pi}$.

**step 3 (decision)** If the new estimates are close enough (following a preset criterion) to the current estimates, accept the new estimates as the solution and stop. Otherwise, replace the current estimates by the new estimates and go to step 2.

We give some more detail on how to compute the conditional probability mentioned in the description of the E-step. From Bayes' theorem we can write

$$\widetilde{P}(C = c|s) = \frac{\widetilde{P}(C = c \text{ and } s)}{\widetilde{P}(s)} = \frac{\widetilde{P}(s|C = c)\widetilde{P}(C = c)}{\widetilde{P}(s)}. \tag{50}$$

Using (50) and the fact that $\widetilde{P}(C = 1) = \widetilde{\pi}$, we find that

$$\widetilde{r}_{12}(s) = \frac{\widetilde{P}(C = 1|s)}{\widetilde{P}(C = 2|s)} = \frac{\widetilde{P}(s|C = 1)\widetilde{\pi}}{\widetilde{P}(s|C = 2)(1 - \widetilde{\pi})}. \tag{51}$$

and $\widetilde{P}(s|C = c)$ is given by (41) using the current parameters $\widetilde{\alpha}_c$ and $\widetilde{\beta}_c$, $(c = 1, 2)$. This gives as a result:

$$\widetilde{r}_{12}(s) = \frac{\widetilde{\beta}_1^{\widetilde{\alpha}_1}(\delta + \widetilde{\beta}_2)^{s+\widetilde{\alpha}_2}}{\widetilde{\beta}_2^{\widetilde{\alpha}_2}(\delta + \widetilde{\beta}_1)^{s+\widetilde{\alpha}_1}} \times \prod_{i=0}^{s-1} \frac{\widetilde{\alpha}_1 + i}{\widetilde{\alpha}_2 + i}. \tag{52}$$

Then it is simple to show that

$$\widetilde{P}(C = 1|s) = \frac{\widetilde{r}_{12}(s)}{1 + \widetilde{r}_{12}(s)}. \tag{53}$$

The M-step consists of three separate procedures. The new estimate of $\pi$ is very simple:

$$\widehat{\pi} = \frac{1}{n} \sum_s n_s \widetilde{P}(C = 1|s) \tag{54}$$

where $n = \sum_s n_s$ is the total sample size. In the other two procedures, the parameters of the gamma distribution in each class are estimated, separately for each class. The estimation procedure for a class is carried out in the same way as in the model with a single distribution, with the only difference that in each class $c$ the observed frequencies $n_s$ are replaced by $n_s\widetilde{P}(C = c|s)$.

The positive thing to mention about the EM-algorithm, when carried out properly is that in each iteration the likelihood will increase and that eventually the procedure will converge, irrespective of how strict the criterion

is. A less pleasant feature of the algorithm is that convergence can be very slow such that many iterations are needed until convergence. Moreover, it is not certain that the maximum of the likelihood function is found in this way. This is not due to the EM-algorithm, but to the fact that it is not known whether the likelihood function has a single maximum or several local maxima. Convergence to a local maximum may occur if the initial estimates (see step 0 of the algorithm) are not well chosen. In carrying out the analyses, the algorithm converged several times to a solution with $\widehat{\pi} = 1$, i.e., a solution with a single distribution. By trial and error we found that a starting value $\widetilde{\pi} = 0.5$, two equal $\widetilde{\beta}$-parameters (equal to the initial estimate for the case of a single distribution) and two $\widetilde{\alpha}$-parameters, chosen close to but at either size of the initial estimate of the single distribution case let the algorithm converge to a solution which was certainly acceptable. This acceptability is discussed in the next subsection.

In all analyses the convergence criterion was set to 0.00005 for the parameter $\pi$, i.e. if $|\widetilde{\pi} - \widehat{\pi}| < 0.00005$, the outcome of the last M-step was accepted as the solution.

## 8.3 Validation and norms

To get a clear understanding of the latent class model, the two estimated distributions for M3 are displayed graphically in Figure 7. The estimate of $\pi$ is 0.537, meaning that about 54% of the M3 population belongs to this class. The two inner curves in the Figure represent the distribution in each class, but both curves are scaled such that the total area under the curves correspond to $\pi$ and $(1 - \pi)$ respectively. The outer curve is just the sum of the two inner curves, and the total area under this curve equals one as it should in a probability distribution.

Class 1 is represented by the peaked inner curve, and it is clear that the right tail for values larger than 40, say, is very close to zero. The second class (the flatter one of the two inner curves) has a heavy right tail. The outer curve is the sum of the two inner curves, and we see that in this curve both features of the observed distribution of the scores (see Figure 6), a tight concentration around the mode and a tick right tail are appearing in the

theoretical mixture distribution.

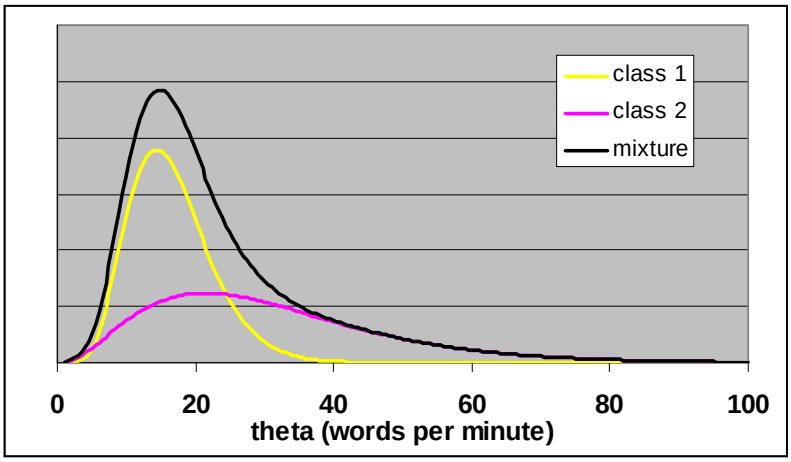


Figure 7. The estimated distributions for M3

Constructing such a curve, however, is a theoretical exercise, and from Figure 7 it can not be concluded that the mixture distribution is valid; in other words, the curves in Figure 7 represent a theoretical construct and one has to check if this construct is in agreement with the observed data.

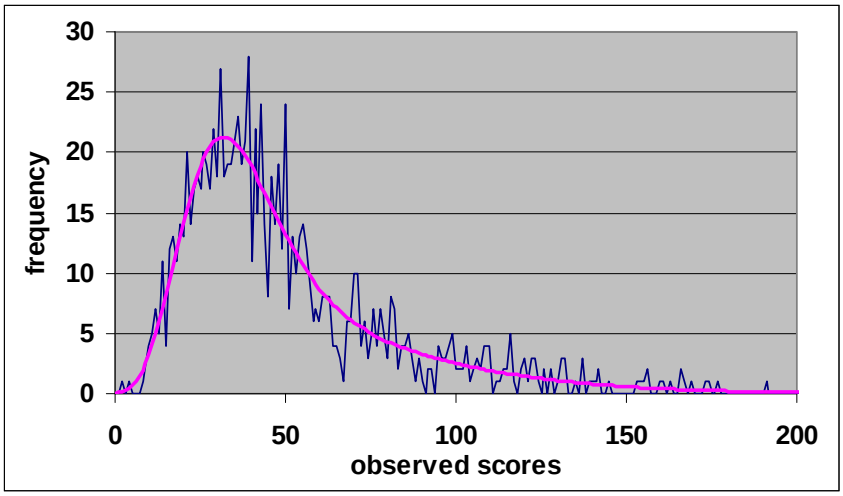


Figure 8. Observed and predicted score distribution for M3

The expected (=predicted) number of students with a score $s$ is given by

$$E(n_s) = n[\pi f_1(s) + (1-\pi) f_2(s)] \tag{55}$$

where $n$ is the sample size and $f_1(s)$ and $f_2(s)$ are given by (41), using the $\alpha$ and $\beta$-parameters for class 1 and 2 respectively. If the theoretical model is valid, then a good correspondence should be found between the expected and observed distribution of the scores. These two distributions are displayed as frequency polygons for M3 in Figure 8. By comparing this figure with Figure 6, a clear improvement is immediately obvious.

To get rid of the irregularities caused by the relative small sample size, cumulative distributions are displayed in Figure 9, jointly for M3 and E3. For E3, a model with two latent classes has been used as well. In both cases a very close agreement between observed and expected frequencies is found, yielding a strong evidence for the validity of the model.

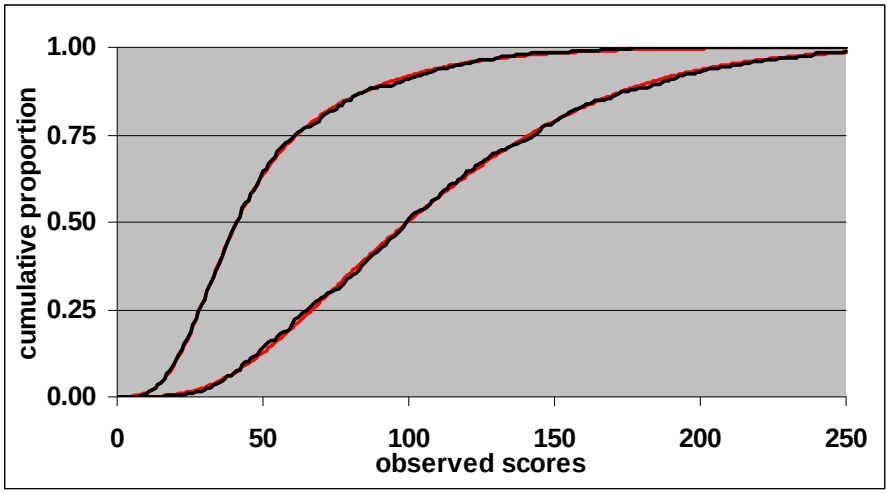


Figure 9. Observed and expected distributions for M3 and E3

To avoid misunderstandings about the meaning of this figure, it should be remembered that students of M3 only got two cards (the two easiest ones), while in E3 each student had to read three cards. The median for M3 is about 41, meaning that the median student from M3 reads about 41 words from the two cards jointly; the median of E3 is about 100, but this is the median score for the three cards jointly. The big shift between the two pairs of curves is the combined effect of being a better reader in E3 and having had three cards instead of two. This figure certainly cannot be used for deriving norm tables. It is only meant as evidence for the validity of the model.

For all grades the latent class model has been used for the TMT, and the correspondence between observed and expected frequencies was very similar to the ones displayed in Figure 9.

Norms are derived from the cumulative distributions of the $\theta$- variable. These distributions are displayed graphically in Figure 10, the dark curves representing from left to right M3, M4,...,M8 and the grey ones (red on a color display) E3, E4,...,E7. The left-most curve (for M3) is the cumulative form of the outer curve in Figure 7.

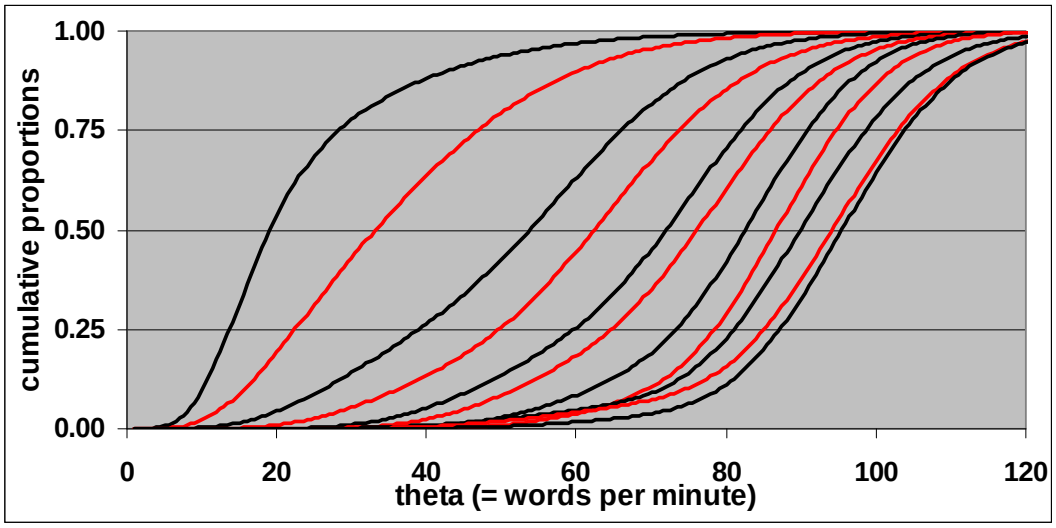


Figure 10. Cumulative distributions of the reading speed for TMT (M3 to M8)

We notice three remarkable features in this figure. At the median (the 0.50 grid line) it is clear that that the largest growth in reading speed occurs in the first year to one and a half year of formal reading instruction (from M3 to M4), the growth from E3 to M4 being larger than in the previous half year. The second remarkable feature is the same as was seen with the tempo tests: from M6 onwards there is scarcely any progress for the weakest 5% of the population.

The most remarkable thing about this figure, however, is the huge variation in M3 and E3, which is mainly due to the best performing quarter of these two populations. A plausible explanation of this phenomenon might be sought in the fact that a number of young students can already read when starting grade 3, either because they have learnt it at home, for example, or because they repeat the third grade.

## 8.4 Further results

### 8.4.1 The $\sigma$-parameters

In Table 4, the estimates of the $\sigma$-parameters are given. Remember that three different cards with different levels of difficulty were constructed (1, 2 and 3) and for each card three parallel forms (labeled a, b and c) were developed. The three parallel forms at each level have very similar parameters estimates, while the three levels clearly differ in difficulty, the smallest value representing the most difficult test. The product of the nine estimates in Table 4 equals one; this is the (arbitrary) normalization.

Table 4. $\sigma$-parameter estimates for the TMT

| level\form | a | b | c | gm |
|---|---|---|---|---|
| 1 | 1.154 | 1.144 | 1.156 | 1.151 |
| 2 | 1.016 | 1.010 | 1.023 | 1.016 |
| 3 | 0.856 | 0.858 | 0.850 | 0.855 |

Since the parallel forms have virtually the same parameter estimates, we will treat them as being really parallel and assign a common value to them. This value is the geometric mean[2] (gm) of the three estimates, and is displayed in the right-most column of the table.

### 8.4.2 The reliability

The use of the latent class model makes the expressions for the reliability a bit more complicated than in the model with a single gamma distribution for reading speed, although the idea is the same: the basic expression is the first equation in (49), which is repeated here for convenience:

$$\rho_{SS\prime} = 1 - \frac{E[Var(S|\theta)]}{Var(S)}.$$

Using (48) gives

$$E[Var(S|\theta)] = \delta E(\theta) = \delta \sum_{c=1}^{2} \pi_c \frac{\alpha_c}{\beta_c} = E(S) \tag{56}$$

[2]The geometric mean of three positive numbers is the cubic root of their product.

where $\pi_1 = \pi$ and $\pi_2 = (1-\pi)$. The parameter $\delta$ in (56) is the parameter for the difficulty of the test. Since we intend to report the reliabilities for each card separately, and $\tau_i = 1$ for each card, the $\delta$-parameter represent just a single $\sigma$-parameter. If cards are combined, the $\delta$-parameter is the sum of the $\sigma$-parameters involved.

For the total variance of the score $S$, we use again the variance decomposition rule, but now conditioning on the latent classes:

$$Var(S) = E[Var(S|C)] + Var[E(S|C)] \tag{57}$$

The first term in the right-hand side of (57) is easily found from (44):

$$E[Var(S|C)] = \sum_{c=1}^{2} \pi_c \frac{\delta\alpha_c}{\beta_c}\left(1 + \frac{\delta}{\beta_c}\right)$$

and using (43) it is readily found that

$$Var[E(S|C)] = \sum_{c=1}^{2} \pi_c \left[\frac{\delta\alpha_c}{\beta_c} - E(S)\right]^2$$

where $E(S)$ is given by (56).

In Table 5 the reliabilities of the cards of the TMT are displayed for all grades. The last column is the reliability for the test consisting of the three cards

Table 5. Reliabilities of the TMT

| grade | 1 | 2 | 3 | 1+2+3 |
|---|---|---|---|---|
| M3 | 0.912 | 0.902 | 0.885 | 0.965 |
| E3 | 0.908 | 0.897 | 0.880 | 0.963 |
| M4 | 0.888 | 0.875 | 0.854 | 0.954 |
| E4 | 0.861 | 0.845 | 0.821 | 0.942 |
| M5 | 0.824 | 0.805 | 0.776 | 0.925 |
| E5 | 0.803 | 0.782 | 0.752 | 0.915 |
| M6 | 0.741 | 0.717 | 0.680 | 0.883 |
| E6 | 0.698 | 0.671 | 0.632 | 0.859 |
| M7 | 0.754 | 0.730 | 0.695 | 0.890 |
| E7 | 0.753 | 0.729 | 0.694 | 0.889 |
| M8 | 0.693 | 0.666 | 0.627 | 0.856 |

In all grades the reliability diminishes with the level of the card, and this is caused by the fact that the difficulty of the cards increases with level while the allotted reading time is held constant. From grade M6 on the reliability might be deemed unacceptably low, and especially in M8 it looks disappointingly low. The main reason for this drop, however, is to be attributed to the decreasing variance of the reading ability with increasing grades on the one hand, and with the particular feature of the Poisson model in which the measurement error increases with increasing value of the variable, i.e., the better students perform, the higher the measurement error is.

### 8.4.3 Local Independence

With the design used in this research, however, another method is available for estimating the reliabilities. The true score of student $v$ on card $i$ equals $\delta_i\theta_v$, from which it easily follows that the correlation (across students) of the true scores on two different cards equals one, i.e., the cards, considered as separate tests, are congeneric. For two congeneric tests it holds that their intercorrelation equals the square root of the product of their reliabilities.

From the analysis on the calibration data (see Table 4), it followed that the parallel versions of the three cards were indeed parallel, so that it is not really necessary to treat parallel versions as different tests. The data collection design was set up in such a way that all students from E3 onwards took three cards, one at each level of difficulty. So for every grade we can compute the empirical correlation between all pairs of cards, and compare these with the expected value that follows from the model. In Table 6, the predicted correlations (from the model) and the empirical correlations are displayed for the grades E3, M5 and M7.

Table 6. Predicted and observed correlations between cards

| | predicted | | | observed | | |
|---|---|---|---|---|---|---|
| pair: | $(1,2)$ | $(1,3)$ | $(2,3)$ | $(1,2)$ | $(1,3)$ | $(2,3)$ |
| E3 | 0.902 | 0.894 | 0.890 | 0.928 | 0.897 | 0.945 |
| M5 | 0.814 | 0.800 | 0.790 | 0.897 | 0.852 | 0.894 |
| M7 | 0.742 | 0.724 | 0.712 | 0.866 | 0.825 | 0.869 |

There are two remarkable and systematic differences between predicted and observed correlations, which also occur in the other grades. The observed correlations are higher, and sometimes substantially higher, than the

predicted ones. The second difference is more subtle, but systematic: for each grade the correlation for the pair $(1, 3)$ is the middle one in the predicted cells, but the lowest one in the observed cells.

The reliability of the test consisting of the three cards jointly can be estimated by using Cronbach's alpha, for example. In Table 7, these estimates are displayed together with the estimates issuing from the Poisson-Gamma model (see right-most column of Table 5).

Table 7. Estimated reliabilities

| | Poisson-Gamma | Cronbach's alpha |
|---|---|---|
| E3 | 0.963 | 0.964 |
| M5 | 0.925 | 0.952 |
| M7 | 0.890 | 0.948 |

The estimates based on the empirical correlations are substantially higher than the ones predicted from the Poisson-Gamma model for the grades M5 and M7.

From these differences, two questions follow, a theoretical one and a practical one. The theoretical question is how to explain this difference, and the practical one is which estimates to use in practical applications. We start with the theoretical problem.

Since the observed correlations are higher than predicted from the Poisson-Gamma model, this model does not incorporate a source of common variation in the data. A basic assumption to most IRT-models is that of local independence, saying essentially that on the three cards the result is driven by a single variable (technical reading ability), and that all deviations in the observed scores from their expected values are independent across cards. This assumption is the same as the assumption of uncorrelated measurement errors in Classical Test Theory. But all data from a single student were collected in a single session, and cards were always presented in the order 1, 2, 3[3]. The general pattern of the correlations (Table 6), shows that the highest correlations are always found in pairs containing the middle card 2. This suggest an autoregressive model of order one, given by

$$\begin{aligned} s_1 &= P(\theta\delta_1) \\ s_t &= (1-\lambda)P(\theta\delta_t) + \lambda s_{t-1}, \quad (t = 2, 3;\ 0 \leq \lambda < 1) \end{aligned} \tag{58}$$

[3]One could object to this invariant order as being a methodological negligence, but in real applications this same order is always maintained, and applying different orders might confuse students, and lead to even more unexplained sources of variance.

where $s_t$ is the observed score on card $t$ and $P(\theta\delta_t)$ is a deviate in a Poisson distribution with parameter $\theta\delta_t$. If the parameter $\lambda = 0$, the model as we have used it results. If $\lambda > 0$, there is a direct influence from the preceding performance. This could be interpreted as the effect of frustration and motivation: a low performance on a card has a negative influence on the following card, and a high performance has a positive influence.

To check if such a simple model could explain the pattern of the correlations as displayed in Table 5, a small simulation study was run. For each of the three grades (E3, M5 and M7) the five parameters of the latent class model were fixed at the value of their estimates in the calibration study and for the three cards the $\delta$-parameters were fixed at the values in the right-most column of Table 4. For each of the three grades, a sample of 1000 students was drawn from the mixture gamma distribution, and for each drawn value of $\theta$, model (58) was applied. For each data set of 3000 students (1000 students in each of the three grades) all parameter values were estimated, and the correlations between the scores on the three cards were computed (per grade) and stored. This whole procedure constitutes a single replication of the simulation study.

Five values of the $\lambda$-parameter were used: 0, 0.05, 0.10, 0.15 and 0.20, and for each value of $\lambda$ twenty replications were carried out.

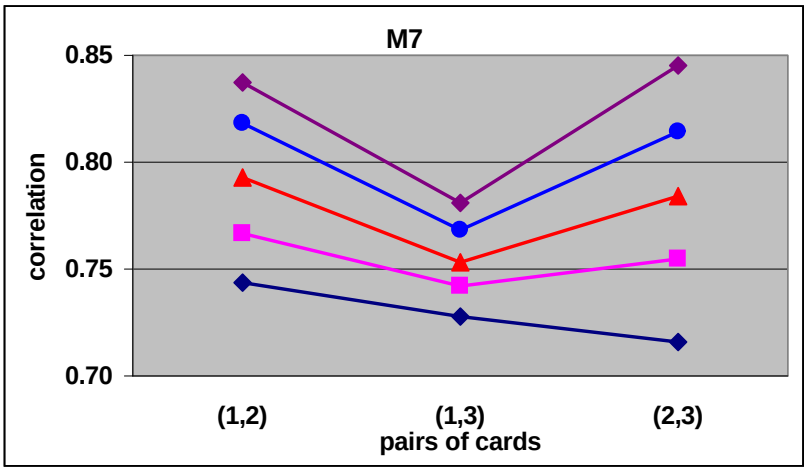


Figure 11. Correlations between scores in simulated data

In Figure 11, the results for the correlations between scores on the three cards are displayed graphically for grade M7. The results for the other two

grades show similar patterns. In the figure, each symbol represents an average correlation over 20 replications. The lowest line corresponds to a $\lambda$ of zero, and lines lie higher with increasing value of $\lambda$. One sees, indeed that even with a quite low positive value of $\lambda$, the typical pattern, visible in the right-hand panel of Table 6 crops up: the two pairs of cards containing card 2 have the highest correlations. Therefore we can conclude that the simple model (58) offers a good explanation of the differences found in Tables 6 and 7.

But as we have solved one problem, we may have created other ones. All conclusions on norms have been based on the validity of the Poisson-Gamma model - implying local independence - but from the results on the correlations and of the simulation study, it is clearly demonstrated that the model is not valid, because the assumption of local independence is violated. Strictly speaking, all conclusions based on the model assumption are invalid, but sticking to such an ultra-othodox point of view would make all work with formal models void, because all assumptions of all models are violated to some degree. A more realistic approach is to investigate to which extent a violation of one or more assumptions influences the inferences made on the results of an analysis using an invalid model. As the main application for the present report is to construct norm tables, i.e. to estimate the distribution of the latent variable, based on all available information, the main criterion to judge the usefulness of the model resides in the correspondence between predicted and observed score distributions. In the simulation study, this correspondence has been checked graphically for the first replication in all three grades and for the five values of $\lambda$, and in all cases the produced figures show an excellent fit as exemplified in Figure 9, which is reassuring.

In summary then, we can state the following conclusions:

1. For each grade (M3 to M8) three cards (two for M3) have been administered to a sample of students (see Table 1 for sample sizes) in a fixed sequence from easy to difficult.

2. From the analysis, it appears that the parallel cards are indeed parallel (see Table 4).

3. From the analysis using the Poisson model as measurement model and the latent class model (a mixture of two gamma distributions) as population model, an excellent correspondence is found between the observed and predicted distribution of the **sum of the scores on all**

**administered tests.** The scores on the separate tests (cards) have only been used to estimate the $\sigma$-parameters.

4. From the analysis of the intercorrelations between the scores on the three cards, it appears clearly that the assumption of local independence is violated. A simple autoregressive model of order 1 can reproduce the pattern of the correlations.

Table 8. Cronbach's alpha

| grade | 1+2 | 1+2+3 |
|---|---|---|
| M3 | 0.964 | – |
| E3 | 0.967 | 0.971 |
| M4 | 0.956 | 0.968 |
| E4 | 0.959 | 0.970 |
| M5 | 0.941 | 0.957 |
| E5 | 0.943 | 0.958 |
| M6 | 0.921 | 0.942 |
| E6 | 0.932 | 0.943 |
| M7 | 0.930 | 0.948 |
| E7 | 0.943 | 0.955 |
| M8 | 0.941 | 0.947 |

As to the practical question on how to evaluate the reliability of the scores, it is clear from the previous analysis that the mere application of the Poisson-Gamma model leads to an underestimation of the reliability, because the assumption of local independence is not fulfilled: correlations between performances on separate cards are systematically higher than predicted by the model, because there are extra sources of covariation beyond the mere technical reading ability. But these sources are systematic and therefore they will contribute to the true score variance. This is the reason why a reliability estimate based on the observed correlations is to be preferred to the theoretical predictions as given in Table 5. A drawback of this approach is that it is not possible, for example, to give a good estimate of the reliability of the cards 2 and 3 separately, because in the data collection design, no students have answered to these cards in isolation. For the same reason, we cannot estimate the reliability of the total test score if the cards would be administered in the reverse order as they have been. What we can do is

to give an estimate for the scores obtained on cards 1 and 2 (administered in that order) and on the cards 1, 2 and 3 administered as they have been. As an estimate of the reliability we use Cronbach 's alpha. The results are displayed in Table 8. The second column is an estimate of the reliability of the sum score on cards 1 and 2; the right-most column is the reliability of the sum score on all three cards.

# 9 Discussion

In this report a psychometric analysis has been applied to two speed tests for technical reading, developed in the framework of Cito's student monitoring system. In both tests, the basic observation is a count. In the tempo tests, the number of correctly completed subtasks is counted; in the three minute test the count is the number of words read correctly within three minutes. The allotted reading time is fixed for all students, but may vary from task to task.

It may be discussed, if the count should reflect the number of subtasks completed or the number of subtasks completed without error. Although this problem is certainly relevant with respect to the construct validity of the test, from a psychometric point of view, the distinction is barely relevant. In fact, for both tests discussed in this report, analyses have been carried out for both cases, and the success in predicting the observed distributions did not show any substantial difference. Of course, the results did differ, by definition one could say, since the number of correct subtasks cannot exceed the number of completed subtasks.

The psychometric model used was originally proposed by Rasch, almost fifty years ago. Rasch, however, was averse to modeling the distribution of the abilities in populations, and considered the work of the psychometrician as finished when reasonable estimates of person abilities could be produced. So for the work reported here, only the part referring to the Poisson model is due to Rasch.

The extension of this model to the Poisson-Gamma model is mainly the work of Jansen and her colleagues at the university of Groningen (e.g., Jansen and van Duijn, 1992; Van Duijn and Jansen, 1995; Jansen, 1997). The detailed elaboration of this model to incomplete designs and the extension to the case of two latent classes is new, and has not been published before. Also new is the proof that JML- and CML-estimates are identical.

One problem with the model, however, remains unsolved. Jansen and Van Duijn (1992) have shown that estimating the $\sigma$-parameters by CML or estimating them jointly with the parameters of the gamma distribution (by marginal maximum likelihood, MML) leads to identical results in a complete design, i.e., a design where all students have taken the same set of tasks. Their method of proof, however, does not generalize to the case of incomplete designs (as the ones that were used for this report). If different estimates would result, this would have an impact on the estimates of the gamma parameters as well. The impact will probably not be very important since the CML-estimates are consistent and the sample sizes used were not very small, but from an academic point of view, it would be reassuring if the differences could be evaluated.

The most important finding, however, in preparing this report is the result that the model has a practical value in a large scale testing system, and this finding outweighs greatly the remaining psychometric and statistical problems.

# References


Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, series B*, 39:1–38.

Jansen, M. (1997). Rasch's model for reading speed with manifest explanatory variables. *Psychometrika*, 62:393–409.

Jansen, M. and van Duijn, M. (1992). Extensions of Rasch's multiplicative Poisson model. *Psychometrika*, 57:405–414.

Lord, F. and Novick, M. (1968). *Statistical Theories of Mental Test Scores*. Reading, Massachussets: Addison-Wesley.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press.

Staphorsius, G. (1994). *Leesbaarheid en Leesvaardigheid*. PhD thesis, University of Twente.

Van Duijn, M. and Jansen, M. (1995). Modeling repeated count data: some extensions of the Rasch Poisson counts model. *Journal of Educational and Behavioral Statistics*, 20:241–258.