Statistics
in Medicine

Research Article

A log-linear multidimensional Rasch model for capture–recapture

E. Pelle,[a*†] D. J. Hessen[b] and P. G. M. van der Heijden[b,c]

In this paper, a log-linear multidimensional Rasch model is proposed for capture–recapture analysis of registration data. In the model, heterogeneity of capture probabilities is taken into account, and registrations are viewed as dichotomously scored indicators of one or more latent variables that can account for correlations among registrations. It is shown how the probability of a generic capture profile is expressed under the log-linear multidimensional Rasch model and how the parameters of the traditional log-linear model are derived from those of the log-linear multidimensional Rasch model. Finally, an application of the model to neural tube defects data is presented. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Rasch model; capture–recapture; heterogeneity; log-linear model; measurement invariance; EM algorithm

1. Introduction

Capture–recapture methods are statistical procedures originally used to estimate the size of wildlife populations [1]. Such procedures are based on a sequence of trapping experiments where individual trapping histories are used to estimate the population size.

Capture–recapture methods have also been successfully applied to estimate the size of human populations. In the case of a human population, the methods are also referred to as *multiple-recapture*, *multiple-records systems*, and *multiple-records systems methods* [2]. In general, capture–recapture methods can be applied in any situation in which two or more incomplete but overlapping lists or registrations are available. Each such registration is then regarded as a capture sample, and the data are usually arranged in an incomplete 2^s contingency table where the missing cell corresponds to the absence in all s registrations. Subsequently, the contingency table is typically analyzed using a log-linear model [3].

Traditional capture–recapture methods assume that the probabilities of inclusion in the s registrations are independent. If dependencies are allowed between registrations, then interaction terms should be included in the log-linear model used [4].

Another assumption in traditional capture–recapture methods is the homogeneity of the capture probability. However, differences of character or behavior between individuals may cause indirect dependence between registrations. Models that were successfully applied to estimate the size of animal and human populations [5, 6] while accounting for unequal catchability are psychometric models, such as the Rasch model.

The Rasch model is a well-known psychometric model for the analysis of dichotomously scored items. In this model, the probability of a response of an individual to an item is modeled as a function of the difficulty of the item and the underlying latent ability of the individual. The use of the dichotomous Rasch model in a capture–recapture context where all registrations are of the same kind provides the possibility to deal with a specific type of heterogeneity, that is, constant apparent dependence between registrations [2, 5, 6].

[a]*Department of Economics, Statistics and Finance, University of Calabria, Arcavacata di Rende, CS, Italy*
[b]*Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands*
[c]*S3RI, University of Southampton, Southampton, U.K.*
*Correspondence to: E. Pelle, Department of Economics, Statistics and Finance, University of Calabria, Arcavacata di Rende, CS, Italy.
†E-mail: elvirapelle@libero.it

An extension of the dichotomous Rasch model is the multidimensional Rasch model, where one or more latent abilities are underlying the responses to the items [7–9]. Bartolucci and Forcina [10] proposed a Rasch-type model for the analysis of capture–recapture data allowing for conditional dependence and multidimensionality; Bartolucci and Pennoni [11] proposed an extension of the latent class model for behavior effects. In the present paper, however, a log-linear multidimensional Rasch model is proposed for capture–recapture analysis of registration data. In the model, registrations are viewed as indicators of a number of latent variables that account for the covariances among registrations. The model is a special case of a log-linear multidimensional partial credit model [12], which was derived using an extension of the Dutch Identity [13].

The paper is organized as follows. First, the log-linear multidimensional Rasch model is discussed in the capture–recapture context. Next, it is shown how the probability of a generic capture profile can be expressed in terms of the log-linear multidimensional Rasch model either with or without a stratifying variable. Subsequently, the connection between the parameters of the log-linear multidimensional Rasch model and those of the standard log-linear model is discussed. Then, the log-likelihood function and the EM algorithm for parameter estimation are described. Finally, the use of the proposed model is illustrated by an application to neural tube defects (NTDs) data.

2. Multidimensional Rasch model for capture–recapture

In this section, we describe the method we propose. First, we deal with the simple situation in which three registrations are available; then, we treat the presence of a stratifying variable. At the end, the extension to a more general situation is discussed.

2.1. Model with three registrations and two latent variables

Consider a situation of three registrations. Let I_1, I_2, I_3 be random variables with respective realizations i_1, i_2, i_3, where $i_s = 0$ if a randomly selected individual is not in registration s and $i_s = 1$ if a randomly selected individual is in registration s, for $s = 1, 2, 3$. Let $n_{i_1 i_2 i_3}$ denote the observed frequency of capture profile $\mathbf{i} = (i_1, i_2, i_3)'$, so that n_{100} denotes the frequency of individuals observed in registration 1 only, n_{110} is the frequency of those observed in registrations 1 and 2 but not in registration 3, and so on. Note that n_{000} is the frequency of individuals not in any registration and has to be estimated in order to estimate the total unknown population size N. The data can be arranged in an incomplete 2^3 contingency table where the missing cell corresponds to the absence in all three registrations (as shown in Table I). To obtain an estimate of n_{000}, we first fit a log-linear model on the incomplete contingency table without the missing cell; then, the parameter estimates of the fitted model are used to predict the value of that part of the population that is missed by all registrations.

Suppose that there are two latent variables that explain the covariances between the registrations. Let $\mathbf{\Theta} = (\Theta_1, \Theta_2)'$ denote the vector of latent variables, and let $\theta = (\theta_1, \theta_2)'$ denote a realization. If the covariances between the random variables I_1, I_2, and I_3 can be explained by θ_1 and θ_2, then I_1, I_2, and I_3 are conditionally independent given the two latent variables. Assume, for example, that registrations 1 and 2 are indicators of the first latent variable and that registrations 2 and 3 are indicators of the second latent variable. Then, a visual presentation of this situation can be given by the path diagram in Figure 1.

In Figure 1, the single-headed arrows from the latent variables to the registrations indicate that there is a direct effect of the latent variables on these registrations, while the curved line between the two

	Table I. Contingency table for three lists.			
		i_3		
	1		0	
	i_2		i_2	
i_1	1	0	1	0
1	n_{111}	n_{101}	n_{110}	n_{100}
0	n_{011}	n_{001}	n_{010}	0^*

*The missing cell is treated as a structurally zero cell.

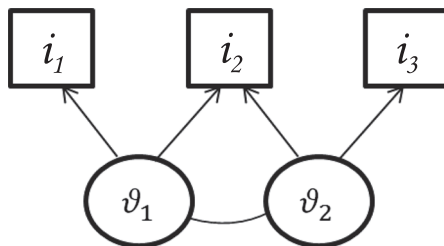E. PELLE, D. J. HESSEN AND P. G. M. VAN DER HEIJDEN

Figure 1. Three registrations and two latent variables.

latent variables indicates that there is a covariance between the two latent variables. On the other hand, as there are no double-headed arrows between pairs of registrations, the registrations are conditionally independent given the latent variables.

It is assumed that the probability of inclusion in registration s given the vector of latent variables equals

$$\pi_{1_s|\theta} = \frac{e^{\mathbf{u}_s'\theta - \delta_s}}{1 + e^{\mathbf{u}_s'\theta - \delta_s}}, \tag{1}$$

where δ_s is the parameter for registration s and $\mathbf{u}_s'$ is the sth row vector of the (3×2) full column rank matrix $\mathbf{U} = [u_{sr}]$ of preassigned weights, where

$$u_{sr} = \begin{cases} 1, & \text{if registration } s \text{ is assumed to be an indicator of latent variable } r, \\ 0, & \text{otherwise.} \end{cases}$$

Note that the probability in Equation (1) is identical to the probability of a positive or correct item score as a function of two latent variables in the multidimensional random effects Rasch model [14], where the latent variables are considered random. For the preceding example, the matrix $\mathbf{U}$ is then given by

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

According to standard probability theory, the probability of a generic capture profile may be written as

$$\pi_{i_1 i_2 i_3} = \int \cdots \int \pi_{i_1 i_2 i_3|\theta} f(\theta) \, d\theta, \tag{2}$$

where $\pi_{i_1 i_2 i_3|\theta}$ is the probability of a generic capture-profile conditional on θ and $f(\theta)$ is the multivariate density of θ in the population of individuals. In the proposed approach, $f(\theta)$ can be left unspecified, as will become clear in what follows. From conditional independence, that is, $\pi_{i_1 i_2 i_3|\theta} = \prod_{s=1}^{3} \pi_{i_s|\theta}$, where $\pi_{i_s|\theta} = (\pi_{1_s|\theta})^{i_s} (1 - \pi_{1_s|\theta})^{1-i_s}$, it now follows that

$$\pi_{i_1 i_2 i_3} = \int \cdots \int \left\{ \prod_{s=1}^{3} \frac{e^{i_s(\mathbf{u}_s'\theta - \delta_s)}}{1 + e^{\mathbf{u}_s'\theta - \delta_s}} \right\} f(\theta) \, d\theta. \tag{3}$$

As $\pi_{000|\theta} = 1 / \left\{ \prod_{s=1}^{3} \left(1 + e^{\mathbf{u}_s'\theta - \delta_s} \right) \right\}$ and according to Bayes's theorem, the posterior distribution of θ given the capture pattern $(i_1, i_2, i_3) = (0, 0, 0)$ equals $g(\theta|0, 0, 0) = \pi_{000|\theta} f(\theta) / \pi_{000}$, and it follows that

$$\pi_{i_1 i_2 i_3} = \pi_{000} e^{-\sum_s i_s \delta_s} \int \cdots \int e^{\mathbf{t}'\theta} g(\theta|0, 0, 0) \, d\theta, \tag{4}$$

where $\mathbf{t} = \mathbf{U}'\mathbf{i}$ and $g(\theta|0, 0, 0)$ is the multivariate density of θ in that part of the population that is not observed in any registration.

Note that

$$M_{\Theta}(\mathbf{t}) = \int \cdots \int e^{\mathbf{t}'\theta} g(\theta|0,0,0)\, d\theta$$

is the moment-generating function conditional on the capture pattern $(i_1, i_2, i_3) = (0, 0, 0)$. In order to compute the probability in Equation (4), it is necessary to make an assumption about the posterior distribution of the latent variables given capture pattern $(i_1, i_2, i_3) = (0, 0, 0)$ and thus to choose a moment-generating function. Here, it is assumed that this posterior distribution follows a multivariate normal distribution, so that

$$M_{\Theta}(\mathbf{t}) = e^{\mathbf{t}'\mu + \frac{1}{2}\mathbf{t}'\Gamma\mathbf{t}}, \tag{5}$$

where μ is the mean vector of θ conditional on capture pattern $(i_1, i_2, i_3) = (0, 0, 0)$ and Γ is the covariance matrix of θ conditional on capture pattern $(i_1, i_2, i_3) = (0, 0, 0)$. Then, the probability of a generic capture profile $\pi_{i_1 i_2 i_3}$ can be expressed as

$$\pi_{i_1 i_2 i_3} = \pi_{000} \exp\left(\sum_{s=1}^{3} i_s \delta_s + t_1 \mu_1 + t_2 \mu_2 + \frac{1}{2} t_1^2 \gamma_{11} + \frac{1}{2} t_2^2 \gamma_{22} + t_1 t_2 \gamma_{12} \right)$$

$$= \pi_{000} \exp\left(\sum_{s=1}^{3} i_s \delta_s + \mathbf{t}'\mu + \frac{1}{2}\mathbf{t}'\Gamma\mathbf{t} \right), \tag{6}$$

where $\mathbf{t} = (t_1, t_2)' = \mathbf{i}'\mathbf{U}$, $\mu = (\mu_1, \mu_2)'$, and $\Gamma = [\gamma_{ir}]$ is symmetric.

Let n be the total number of individuals observed in at least one registration. Let $A = \{(1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)\}$ be the set of capture profiles of individuals observed in at least one registration. As the observed frequencies $n_{100}, n_{010}, n_{001}, n_{110}, n_{101}, n_{011}, n_{111}$ have a multinomial distribution with parameters n and $\pi_{i_1 i_2 i_3} / \sum_A \pi_{i_1 i_2 i_3}$, for all $(i_1, i_2, i_3) \in A$, we can express the expected frequency of $n_{i_1 i_2 i_3}$ as

$$m_{i_1 i_2 i_3} = n\pi_{i_1 i_2 i_3} / \sum_A \pi_{i_1 i_2 i_3}, \text{ for all } (i_1, i_2, i_3) \in A. \tag{7}$$

Substituting from Equation (6) into Equation (7) and taking the logarithm yields the log-linear representation

$$\ln m_{i_1 i_2 i_3} = \delta + \sum_{s=1}^{3} i_s \delta_s + \mathbf{t}'\mu + \frac{1}{2}\mathbf{t}'\Gamma\mathbf{t}, \tag{8}$$

where $\delta = \ln\left(n\pi_{000} / \sum_A \pi_{i_1 i_2 i_3} \right) = \ln\left\{ n / \sum_A \exp\left(\sum_{s=1}^{3} i_s \delta_s + \mathbf{t}'\mu + \frac{1}{2}\mathbf{t}'\Gamma\mathbf{t} \right) \right\}$.

Note that the model in Equation (8) is not identified. Because of the specific choice of $\mathbf{U}$, $t_1 = u_{11}i_1 + u_{21}i_2 + u_{31}i_3 = i_1 + i_2$ and $t_2 = u_{12}i_1 + u_{22}i_2 + u_{32}i_3 = i_2 + i_3$, so that $\mathbf{t}'\mu = t_1 \mu_1 + t_2 \mu_2 = i_1 \mu_1 + i_2(\mu_1 + \mu_2) + i_3 \mu_2$. Consequently, $\sum_{s=1}^{3} i_s \delta_s + \mathbf{t}'\mu$ in Equation (8) equals $i_1(\delta_1 + \mu_1) + i_2(\delta_2 + \mu_1 + \mu_2) + i_3(\delta_3 + \mu_2)$, and δ_1, δ_2, and δ_3 cannot be separated from μ_1 and/or μ_2. To go around this problem, we arbitrarily fix μ to $\mathbf{0}$. Then, the model can be rewritten as

$$\ln m_{i_1 i_2 i_3} = \delta + \sum_{s=1}^{3} i_s \delta_s + \frac{1}{2}\mathbf{t}'\Gamma\mathbf{t}$$

$$= \delta + i_1 \delta_1 + i_2 \delta_2 + i_3 \delta_3 + \frac{1}{2} t_1^2 \gamma_{11} + \frac{1}{2} t_2^2 \gamma_{22} + t_1 t_2 \gamma_{12}, \tag{9}$$

where δ is a common effect parameter, δ_s is the main-effect parameter for registration s, γ_{11} is the variance of the first latent variable given t_1 and t_2, γ_{22} is the variance of the second latent variable given t_1 and t_2, and γ_{12} is the covariance between the two latent variables given t_1 and t_2. For convenience, the resulting model will be denoted by $i_1 + i_2 + i_3 + t_1 + t_2$. Note that there are $2(2 + 1)/2 = 3$ parameters to account for the two latent variables θ_1 and θ_2.

Let $\mathbf{m} = (m_{100}, m_{010}, m_{001}, m_{110}, m_{101}, m_{011}, m_{111})'$ be the vector of expected counts. In matrix terms, the model in Equation (9) may be written as $\ln \mathbf{m} = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\delta, \delta_1, \delta_2, \delta_3, \gamma_{11}, \gamma_{22}, \gamma_{12})'$ is the vector of parameters to be estimated and $\mathbf{X}$ is the design matrix with columns corresponding to the parameters to be estimated, that is, $\mathbf{X} = (\mathbf{1}, \mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3, \mathbf{t}_1^2, \mathbf{t}_2^2, \mathbf{t}_1\mathbf{t}_2)'$. If we suppose that registrations 1 and 2 are indicators of the first latent variable and that registrations 2 and 3 are indicators of the second latent variable, then matrix $\mathbf{X}$ may be written as

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 4 & 1 & 2 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 4 & 2 \\ 1 & 1 & 1 & 1 & 4 & 4 & 4 \end{pmatrix},$$

and the model may be fitted as a traditional log-linear model, hence the name log-linear multidimensional Rasch model.

2.2. Model with three registrations, two strata, and two latent variables

Suppose now that registrations are recorded in two strata (or time periods, for example, 2 years). In this situation, year is a stratifying variable with two categories denoted by the index j and $n_{i_1 i_2 i_3 j}$, and $\pi_{i_1 i_2 i_3 j}$ denote the observed frequency and the probability for year j, respectively.

The resulting contingency table has two missing cells corresponding to the capture profile of not being observed in any registration for each stratum (as shown in Table II). **T2**

The probability of a generic capture profile may be written as

$$\pi_{i_1 i_2 i_3 j} = \int \cdots \int \pi_{i_1 i_2 i_3 j | \boldsymbol{\theta}} f(\boldsymbol{\theta}) \, d\boldsymbol{\theta}, \tag{10}$$

where $\pi_{i_1 i_2 i_3 j | \boldsymbol{\theta}}$ is the probability of capture profile (i_1, i_2, i_3) for year j conditional on the vector of latent variables and $f(\boldsymbol{\theta})$ is the multivariate density of $\boldsymbol{\theta}$.

Assuming that the posterior distribution of the latent variables given the capture pattern $(i_1, i_2, i_3) = (0, 0, 0)$ follows a multivariate normal distribution, the model in Equation (6) can be written as

$$\pi_{i_1 i_2 i_3 j} = \pi_{000j} \exp\left(\sum_{s=1}^{3} i_s \delta_{sj} + \mathbf{t}'\boldsymbol{\mu}_j + \tfrac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}_j\mathbf{t} \right), \tag{11}$$

where $\boldsymbol{\mu}_j$ is the mean vector and $\boldsymbol{\Gamma}_j$ is the covariance matrix of $\boldsymbol{\Theta}$ in stratum j.

Let $m_{i_1 i_2 i_3 j}$ denote the expected frequency corresponding to the observed frequency $n_{i_1 i_2 i_3 j}$, that is,

$$m_{i_1 i_2 i_3 j} = n\pi_{i_1 i_2 i_3 j} \Big/ \sum_A \pi_{i_1 i_2 i_3 j}, \text{ for all } (i_1, i_2, i_3) \in A. \tag{12}$$

		i_3			
		\multicolumn{2}{c}{1}	\multicolumn{2}{c}{0}
		i_2		i_2	
Year	i_1	1	0	1	0
1	1	n_{1111}	n_{1011}	n_{1101}	n_{1001}
	0	n_{0111}	n_{0011}	n_{0101}	0^*
2	1	n_{1112}	n_{1012}	n_{1102}	n_{1002}
	0	n_{0112}	n_{0012}	n_{0102}	0^*

Table II. Contingency table for three lists and two strata.

*Missing cells are treated as structurally zero cells.

Thus, we obtain

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^{3} i_s \delta_{sj} + \mathbf{t}' \boldsymbol{\mu}_j + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_j \mathbf{t}, \tag{13}$$

where $\delta_j = \ln(n \pi_{000j} / \sum_A \pi_{i_1 i_2 i_3 j})$. Without any additional constraints, the model in (13) is not identified; setting $\boldsymbol{\mu}_j$ equal to zero for identification, we have

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^{3} i_s \delta_{sj} + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_j \mathbf{t}, \tag{14}$$

where δ_j is the common effect parameter in stratum j and δ_{ij} is the main-effect parameter for registration i in stratum j.

2.2.1. Model of measurement invariance.
Assume now that parameters are equal across strata. This means that the model has measurement invariance across strata (that is, the model applies across strata). Under the assumption of measurement invariance, we have

$$\delta_{sj} = \delta_s, \qquad \forall j. \tag{15}$$

Thus, the model in Equation (13) is equal to

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^{3} i_s \delta_s + \mathbf{t}' \boldsymbol{\mu}_j + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_j \mathbf{t}. \tag{16}$$

Without additional constraints, this model is not identified. Because of the assumption of measurement invariance, we now only need to set $\boldsymbol{\mu}_j$ to $\mathbf{0}$, for one j, to identify the model.

In the case of measurement invariance, it is possible to test whether $\boldsymbol{\mu}_j = \boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}$ for all j. If this simultaneous hypothesis holds, then the model in Equation (16) becomes

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^{3} i_s \delta_s + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma} \mathbf{t}, \tag{17}$$

where the parameters can be interpreted as before.

2.3. General case

The extension of the method described in the preceding sections to a more general situation is straightforward. Assume that we have S registrations and J strata. Let $n_{i_1,\dots,i_S j}$ and $\pi_{i_1,\dots,i_S j}$ be the observed frequencies and the probabilities, respectively, where $j = 1, 2, \dots, J$. Note that the resulting contingency table has J structural zeros (one for each stratum).

Suppose now that the covariances between the random variables $I_1, \dots, I_S$ can be explained by q latent variables. Let $\mathbf{u}'_s$ denote the sth row of the $SJ \times q$ full column rank matrix $U = [u_{sr}]$, where $u_{sr} = 1$ if registration s belongs to the rth latent variable and 0 otherwise, and let $\mathbf{t} = (t_1, \dots, t_q)$ be the vector of the total scores, where $t_r = \sum_{s=1}^{S} u_{sr} i_s$, for $r = 1, \dots, q$.

Under the assumption of a multivariate normal posterior distribution of the latent variables (conditional on the capture pattern of individuals not observed in any registration), the probability of a generic capture profile $\pi_{i_1,\dots,i_S j}$ is equal to

$$\pi_{i_1,\dots,i_S j} = \pi_{0\dots0} \exp\left(\sum_{s=1}^{S} i_s \delta_{sj} + \mathbf{t}' \boldsymbol{\mu}_j + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_j \mathbf{t} \right), \tag{18}$$

where $\boldsymbol{\mu}_j$ is the mean vector for the jth stratum and $\boldsymbol{\Gamma}_j$ is a symmetric matrix. Let $m_{i_1,\dots,i_S j} = n \pi_{i_1,\dots,i_S j}$ denote the expected count of observed frequency $n_{i_1,\dots,i_S j}$. Then, we have the log-linear representation

$$\ln m_{i_1,\dots,i_S j} = \delta_j + \sum_{s=1}^{S} i_s \delta_{sj} + \mathbf{t}' \boldsymbol{\mu}_j + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_j \mathbf{t}. \tag{19}$$

Without additional constraints, the model is not identified. If we set $\boldsymbol{\mu}_j$ equal to $\mathbf{0}$ for identification, then the model becomes

$$\ln m_{i_1,\ldots,i_Sj} = \delta_j + \sum_{s=1}^{S} i_s \delta_{sj} + \tfrac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}_j\mathbf{t}. \tag{20}$$

dient The model in Equation (20) is a traditional log-linear model, and, once the parameters have been estimated, an estimate of the portion of the population missed by all registrations and an estimate of the total unknown population size N can be obtained.

If the assumption of measurement invariance holds, then the model in (19) can be written in the following way:

$$\ln m_{i_1,\ldots,i_Sj} = \delta_j + \sum_{s=1}^{S} i_s \delta_s + \mathbf{t}'\boldsymbol{\mu}_j + \tfrac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}_j\mathbf{t}, \tag{21}$$

where the parameters can be interpreted as before.

3. Connection with the standard log-linear model

In applying the log-linear unidimensional Rasch model to capture–recapture data, a standard log-linear model is assumed in which all two-factor interaction parameters are equal and positive (this model is denoted in [2] as a model with a first-order heterogeneity term H_1, but the link with the log-linear unidimensional Rasch model is not explicitly made). In applying the multidimensional log-linear Rasch model to capture–recapture data, the structure of the two-factor interaction parameters of the corresponding standard log-linear model depends on the specific assumptions about the relationships between the registrations and the latent variables. In the following, a reparameterization is given in which the parameters of the standard log-linear model are expressed in terms of the parameters of the log-linear multidimensional Rasch model.

The standard log-linear model in which all two-factor interaction parameters are present can be written as

$$\ln m_{i_1,\ldots,i_Sj} = \lambda_j + \sum_{s=1}^{S} i_s \lambda_{sj} + \sum_{s=1}^{S-1}\sum_{c=s+1}^{S} i_s i_c \lambda_{scj}, \tag{22}$$

where λ_j denotes a main-effect parameter for stratum j, λ_{sj} denotes a main-effect parameter for the sth registration in the jth stratum, and λ_{scj} denotes a two-factor interaction parameter for registrations c and s in stratum j.

Writing out t_r^2 and $t_r t_v$, after some algebra, the multidimensional Rasch model in Equation (20) takes the following form:

$$\begin{aligned}
\ln m_{i_1,\ldots,i_Sj} = \delta_j &+ \sum_{s=1}^{S} i_s \left[\delta_{sj} + \tfrac{1}{2}\sum_{r=1}^{q} u_{sr}\gamma_{rrj} + \sum_{r=1}^{q-1}\sum_{v=r+1}^{q} u_{sr}u_{sv}\gamma_{rvj} \right] \\
&+ \sum_{s=1}^{S-1}\sum_{c=s+1}^{S} i_s i_c \left[\sum_{r=1}^{q} u_{sr}u_{cr}\gamma_{rrj} + \sum_{r=1}^{q-1}\sum_{v=r+1}^{q} \left(u_{sr}u_{cv} + u_{sv}u_{cr} \right)\gamma_{rvj} \right],
\end{aligned} \tag{23}$$

and it is possible to compute the parameters of the standard log-linear model in Equation (22) starting from those of the multidimensional Rasch model using

$$\lambda_j = \delta_j, \tag{24}$$

and

$$\lambda_{sj} = \delta_{sj} + \tfrac{1}{2}\sum_{r=1}^{q} u_{sr}\gamma_{rrj} + \sum_{r=1}^{q-1}\sum_{v=r+1}^{q} u_{sr}u_{sv}\gamma_{rvj}, \tag{25}$$

and

$$\lambda_{scj} = \sum_{r=1}^{q} u_{sr} u_{cr} \gamma_{rrj} + \sum_{r=1}^{q-1} \sum_{v=r+1}^{q} \left(u_{sr} u_{cv} + u_{sv} u_{cr} \right) \gamma_{rvj}. \tag{26}$$

In other words, through these formulas, the parameters of the standard log-linear model can be computed from the parameters of the log-linear multidimensional Rasch model, regardless of the structure of the latent variables [12].

4. Parameter estimation

The log likelihood for the general case of S registrations and J strata is given by

$$l = \sum_{j=1}^{J} n_j \delta_j + \sum_{j=1}^{J} \sum_{s=1}^{S} n_{sj} \delta_{sj} + \sum_{j=1}^{J} \sum_{\mathbf{t}} n_{\mathbf{t}j} \left(\mathbf{t}' \boldsymbol{\mu}_j + \tfrac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_j \mathbf{t} \right), \tag{27}$$

where n_j is the number of individuals in stratum j, n_{sj} is the number of individuals in stratum j and in registration s, and $n_{\mathbf{t}j}$ is the number of individuals in stratum j with the observed vector of total scores $\mathbf{t}$. Given restrictions needed for one of the special cases discussed previously, the log likelihood can be maximized with respect to the parameters subject to identification constraints.

The data are incomplete because of the unknown frequency n_{000}. However, assuming data are missing at random, the EM algorithm can be used to compute the maximum likelihood estimate of the population size. In particular, in the qth iteration of the E-step, the expected frequencies are calculated, where the expected frequency for n_{000} is derived from the parameter estimates found at iteration $q - 1$. Once all the expected frequencies are computed and the dataset is completed, in the M-step, a log-linear model is fitted to the completed data, and the log likelihood is maximized in order to calculate the probability estimates that will be used in the $(q + 1)$th iteration of the E-step. Thus, the updates for the completed data are derived, and the log-linear model is fitted in the M-step. This procedure is repeated until the log-likelihood function converges. The final parameter estimates are used to estimate the expected frequencies for the structural zero cells, and an estimate of the total population size is calculated.

5. Application

To illustrate the methodology of the preceding sections, the data from the five registrations described by Zwane *et al.* [15] on NTDs in the Netherlands are used. The five registrations on NTDs cover different but overlapping periods of time. Zwane *et al.* [15] showed that if the fact that registrations refer to different but overlapping populations is ignored, then the resulting estimates of the total population size may be biased. They approached this situation as a missing data problem and presented a version of the EM algorithm to estimate the missing entry resulting from registrations that are not operating in some strata. We will use the EM algorithm to analyze the data. All computations were carried out using the statistical ʀ program.

We now motivate the models that we will fit to the data. Model 1 is a classical model that can be used as a baseline. It assumes that the five registrations are independent and adds another set of 10 parameters to allow the sizes of the 11 years to differ. Model 2 expands model 1 by including an interaction parameter for each pair of registrations. As there are five registrations, 10 extra parameters are added. In model 3, is it assumed that these 10 interaction parameters are identical; thus, the number of parameters is reduced with 9. This is the log-linear version of the unidimensional Rasch model, which is also found in [2], and described as a log-linear model with heterogeneity of order 1 (H_1). It is included in our list of models to compare its fit with the fit of multidimensional Rasch models.

To apply the multidimensional Rasch model to the dataset on NTDs in the Netherlands, we assume that the five registrations may be divided into two sets of indicators that each measure a separate latent variable.

In order to decide which registrations measure the same latent variable, we study the parameter estimates of the two-factor interactions of model 2. Table III summarizes the estimates for the two-factor interaction parameters among registrations. A high value of an estimate of a two-factor interaction is an indication of a positive relationship between two registrations. Such registrations can then be viewed as indicators of the same latent variable.

Table III. Estimates of the two-factor interaction parameters.

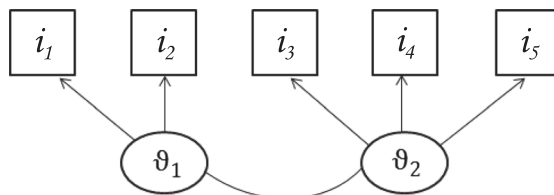|			c		
s	1	2	3	4	5
1	—				
2	0.718424	—			
3	0.185740	0.024525	—		
4	0.557406	1.055780	1.690401	—	
5	0.633640	−0.100489	0.467334	1.725820	—

Figure 2. Model with five registrations and two latent variables.
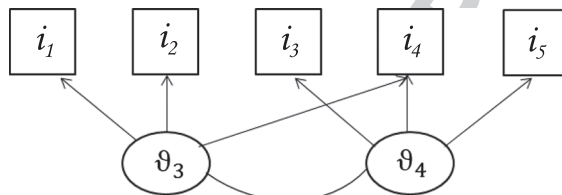
Figure 3. Model with five registrations and two latent variables.

From Table III, it can be concluded that registrations 1 and 2 measure a first latent variable (named θ_1) and that registrations 3, 4, and 5 measure a second latent variable (called θ_2). A visual presentation of this situation is given by the path diagram in Figure 2. Assuming measurement invariance, the model is given by

$$\ln m_{i_1 i_2 i_3 i_4 i_5 j} = \delta + \delta_j + \sum_{s=1}^{5} i_s \delta_s + \frac{1}{2} \sum_{r=1}^{2} t_r^2 \gamma_{rr} + t_1 t_2 \gamma_{12}, \quad \text{for } j = 1988, \ldots, 1997,$$

where $\mathbf{t} = (t_1, t_2)' = \mathbf{i}'\mathbf{U}$ are the total scores accounting for the latent variables θ_1 and θ_2, respectively; δ is the common effect parameter; and δ_j is the main-effect parameter for year j (here. year 1998 was chosen as reference category). For this model, the matrix $\mathbf{U}$ of weights for the latent variables is given by

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \\ u_{41} & u_{42} \\ u_{51} & u_{52} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

and the total scores are $t_1 = i_1 + i_2$ and $t_2 = i_3 + i_4 + i_5$. This is model 4.

From Table III, it seems also reasonable to conclude that registrations 1, 2, and 4 measure the same latent variable (say θ_3) and that registrations 3, 4, and 5 are indicators of the another latent variable (named θ_4). In this case, the two latent variables have registration 4 in common. Figure 3 shows this situation.

Under measurement invariance, the model is now given by

$$\ln m_{i_1 i_2 i_3 i_4 i_5 j} = \delta + \delta_j + \sum_{s=1}^{5} i_s \delta_s + \frac{1}{2} \sum_{r=3}^{4} t_r^2 \gamma_{rr} + t_3 t_4 \gamma_{34}, \quad \text{for } j = 1988, \ldots, 1997.$$

In this model, the matrix $\mathbf{U}$ is given by

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \\ u_{41} & u_{42} \\ u_{51} & u_{52} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix},$$

so that the total scores are $t_3 = i_1 + i_2 + i_4$ and $t_4 = i_3 + i_4 + i_5$. This is model 5.

Table IV summarizes the results of the models fitted to the data. In Table IV(a), for each model, the number of parameters, the degrees of freedom, the deviance, the value of AIC, the value of BIC, and the estimate of the total population size $\hat{N}$ are reported. In Table IV(b), the yearly estimates $\hat{N}_j$, for $j = 1988, \ldots, 1998$, under each model are presented.

In Figure 4, the yearly estimates for each model are plotted.

Model 1, the log-linear model with main-effect parameters and parameters for year, does not fit the data well and has a high deviance. Model 2, the model with a different estimate for the interaction between each

Table IV. Selected models.

(a) Selected models with deviance, AIC and BIC

Model	Design matrix	Par	df*	Dev	AIC	BIC	$\hat{N}$
1	$i_1 + i_2 + i_3 + i_4 + i_5 + Y_{cat}$	16	213	400	432	487	2229
2	$1 + (i_1 i_2 + \cdots + i_4 i_5)$	26	203	298	350	439	3077
3	$1 + H_1$	17	212	349	383	441	3009
4	$1 + t_1 + t_2$	19	210	324	362	427	2793
5	$1 + t_3 + t_4$	19	210	311	349	414	3041

(b) Selected models with yearly estimates

Model	$\hat{N}_{88}$	$\hat{N}_{89}$	$\hat{N}_{90}$	$\hat{N}_{91}$	$\hat{N}_{92}$	$\hat{N}_{93}$	$\hat{N}_{94}$	$\hat{N}_{95}$	$\hat{N}_{96}$	$\hat{N}_{97}$	$\hat{N}_{98}$
1	199	224	234	206	222	186	189	202	178	210	179
2	275	309	323	285	302	258	261	280	246	290	248
3	272	305	319	281	303	249	252	271	238	280	239
4	251	282	295	260	280	232	235	252	222	261	223
5	271	305	318	281	300	255	258	277	244	287	245

There are 229 observed cells. H_1 is the first-order heterogeneity term. $t_1 = i_1 + i_2$ and $t_2 = i_3 + i_4 + i_5$. $t_3 = i_1 + i_2 + i_4$ and $t_4 = i_3 + i_4 + i_5$.
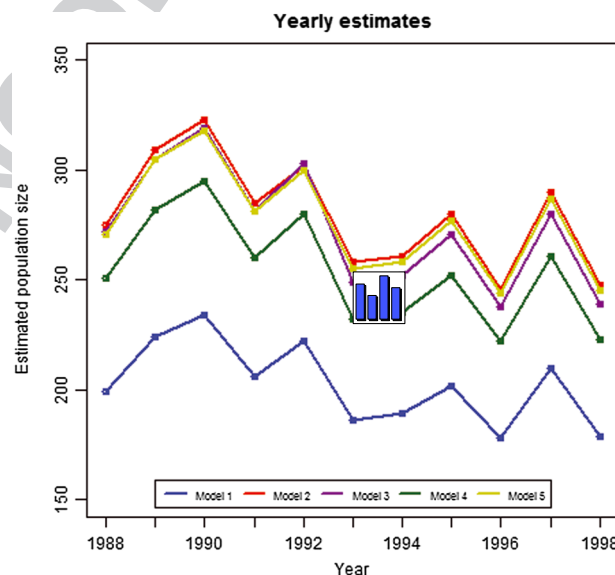

Figure 4. Yearly estimates for the five models.

pair of registrations, has a much better fit in terms of AIC and BIC. Model 3, the unidimensional Rasch model, uses only one parameter more than model 1, and using this single parameter, it accomplishes a fit in between models 1 and 2. Both of the multidimensional Rasch models, models 4 and 5, fit well to the data and have a smaller deviance than the unidimensional Rasch model. Model 5, where registration 4 is an indicator for both latent variables, is the best model because it has the smallest AIC and BIC values. Therefore, this model is selected as the final model.

Table V. Selected models.

Parameter	Estimate	Standard error
(a) Multidimensional Rasch model		
δ	4.513951	0.142557
δ_{1988}	0.101292	0.116082
δ_{1989}	0.218309	0.112754
δ_{1990}	0.260357	0.111628
δ_{1991}	0.135194	0.115088
δ_{1992}	0.201906	0.111082
δ_{1993}	0.038221	0.112887
δ_{1994}	0.050644	0.112545
δ_{1995}	0.122103	0.110637
δ_{1996}	−0.00651	0.114147
δ_{1997}	0.156004	0.109768
δ_1	−2.20858	0.14922
δ_2	−1.04768	0.142911
δ_3	−3.25652	0.124767
δ_4	−2.9981	0.176131
δ_5	−4.16525	0.145811
γ_{33}	0.618927	0.082545
γ_{44}	1.108461	0.087735
γ_{34}	0.219176	0.053513
(b) Log-linear model		
λ_1	−1.89911	0.154823
λ_2	−0.73821	0.148751
λ_3	−2.70229	0.132254
λ_4	−1.91523	0.193684
λ_5	−3.61102	0.152267
λ_{12}	0.618927	0.082545
λ_{13}	0.219176	0.053513
λ_{14}	0.838102	0.098373
λ_{15}	0.219176	0.053513
λ_{23}	0.219176	0.053513
λ_{24}	0.838102	0.098373
λ_{25}	0.219176	0.053513
λ_{34}	1.327637	0.101656
λ_{35}	1.108461	0.087735
λ_{45}	1.327637	0.101656

Table VI. 95% confidence intervals.

Model	Design matrix	$\hat{N}$	95% CI
1	$i_1 + i_2 + i_3 + i_4 + i_5 + Y_{cat}$	2229	[2164, 2297]
2	$1 + (i_1 i_2 + \cdots + i_4 i_5)$	3077	[2724, 3571]
3	$1 + H_1$	3009	[2737, 3345]
4	$1 + t_1 + t_2$	2793	[2559, 3104]
5	$1 + t_3 + t_4$	3041	[2755, 3409]

H_1 is the first-order heterogeneity term. $t_1 = i_1 + i_2$ and $t_2 = i_3 + i_4 + i_5$. $t_3 = i_1 + i_2 + i_4$ and $t_4 = i_3 + i_4 + i_5$.

Equations (25) and (26) can now be used to obtain an expression for the standard log-linear model in terms of the parameters of model 5. Here, Equations (25) and (26) simplify to

$$\lambda_s = \delta_s + \frac{1}{2}\sum_{r=3}^{4} u_{sr}\gamma_{rr} + u_{s3}u_{s4}\gamma_{34},$$

$$\lambda_{sc} = \sum_{r=3}^{4} u_{sr}u_{cr}\gamma_{rr} + \left(u_{s3}u_{c4} + u_{s4}u_{c3}\right)\gamma_{34}.$$

Using these equations, we obtain the following expressions for the parameters of the log-linear model:

$$\lambda_1 = \delta_1 + \frac{1}{2}\gamma_{33} \qquad \lambda_2 = \delta_2 + \frac{1}{2}\gamma_{33} \qquad \lambda_3 = \delta_3 + \frac{1}{2}\gamma_{44}$$
$$\lambda_4 = \delta_4 + \frac{1}{2}\left(\gamma_{33} + \gamma_{44}\right) + \gamma_{34} \qquad \lambda_5 = \delta_5 + \frac{1}{2}\gamma_{44} \qquad \lambda_{12} = \gamma_{33}$$
$$\lambda_{13} = \gamma_{34} \qquad \lambda_{14} = \gamma_{33} + \gamma_{34} \qquad \lambda_{15} = \gamma_{34}$$
$$\lambda_{23} = \gamma_{34} \qquad \lambda_{24} = \gamma_{33} + \gamma_{34} \qquad \lambda_{25} = \gamma_{34}$$
$$\lambda_{34} = \gamma_{44} + \gamma_{34} \qquad \lambda_{35} = \gamma_{44} \qquad \lambda_{45} = \gamma_{44} + \gamma_{34}$$

Thus, the main-effect parameters are equal to the main-effect parameters for model 5 plus half of the variance (given the total scores) of the latent variable for which the registration is an indicator, except for the registration 4, for which it is equal to the main-effect parameter δ_4 plus half of the variance of both latent variables plus the covariance between θ_3 and θ_4, given the total scores. Concerning the two-factor interaction parameters, for those involving registrations that are indicators of different latent variables (that are $\lambda_{13}, \lambda_{15}, \lambda_{23}, \lambda_{25}$) are equal to the covariance (γ_{34}) conditional on the total scores. The two-factor interaction parameters that involve registrations measuring the same latent variable (except those involving registration 4) are equal to the variance (given the total scores) of the corresponding latent variable, while other two-factor interaction parameters ($\lambda_{14}, \lambda_{24}, \lambda_{34},$ and λ_{45}) are equal to the covariance (given the total scores) plus the variance (given the total scores) of the latent variable for which the other registration is assumed to be an indicator. Table V(a) reports the parameter estimates for model 5 and the corresponding standard errors. In Table V(b), the parameter estimates of the corresponding standard log-linear model are reported.

To derive confidence intervals, we do not apply asymptotic methods but apply the parametric bootstrap (compare with that of Zwane and van der Heijden [16]). One reason is that the parametric bootstrap allows for non-symmetric confidence intervals. Second, it is not easy to derive asymptotic methods in the current situation where registrations are not observed in every year. To compute the bootstrapped confidence intervals, we first estimate the probabilities for the completed contingency table under a model, including all the cells that cannot be observed by design. For the first bootstrap sample, a multinomial sample is drawn given these parameters, and the sample is then reformatted to be identical to the observed data. The model is then fitted to the reformatted sample, and the population size is estimated. This is the first parametric bootstrap estimate. We used 500 parametric bootstrap samples and the percentile method to compute the confidence intervals for the population size estimates for each of the five models (Table VI); we also computed confidence intervals for yearly estimates of the population size for models 2 and 5. In this case, confidence intervals for the yearly estimates for model 5 are always smaller than those of

Table VII. 95% confidence intervals for yearly estimates of the population size.							
		Model 2		Model 5		Log-linear		
Year	Observed	$\hat{N}$	95% CI	$\hat{N}$	95% CI	Model	$\hat{N}$	95% CI
1988	145	275	[225, 333]	271	[221, 330]	$i_1i_2 + i_5$	311	[200, 648]
1989	163	309	[256, 385]	305	[255, 372]	$i_1 + i_2i_5$	174	[161, 192]
1990	170	323	[272, 395]	318	[268, 394]	$i_1 + i_2i_5$	177	[168, 189]
1991	150	285	[234, 360]	281	[234, 344]	$i_1i_2 + i_1i_5$	191	[149, 282]
1992	172	302	[251, 367]	300	[254, 362]	$i_1i_2 + i_2i_3 + i_5 + H_1$	782	[326, 2687]
1993	160	258	[211, 311]	255	[211, 305]	$i_1i_2 + i_1i_5 + i_2i_4 + i_3i_4 + i_4i_5$	320	[207, 957]
1994	162	261	[216, 325]	258	[215, 319]	$i_1i_4 + i_1i_5 + i_2i_4 + i_3i_4 + i_4i_5$	232	[197, 293]
1995	174	280	[233, 342]	277	[235, 329]	$i_1i_2 + i_1i_3 + i_2i_3 + i_3i_4 + i_3i_5 + i_4i_5$	206	[188, 231]
1996	153	246	[204, 308]	244	[203, 296]	$i_1i_2 + i_1i_4 + i_2i_4 + i_2i_5 + i_3i_4 + i_4i_5$	317	[220, 583]
1997	180	290	[243, 355]	287	[238, 345]	$i_1i_2 + i_1i_4 + i_1i_5 + i_2i_4 + i_3i_4 + i_3i_5 + i_4i_5$	351	[259, 595]
1998	154	248	[200, 308]	245	[205, 301]	$i_1i_4 + i_2i_3 + i_2i_4 + i_2i_5 + i_3i_4 + i_4i_5$	212	[179, 266]

model 2 as shown in Table VII. Here, yearly estimates of the population size and confidence intervals for the standard log-linear model are presented. Note that traditional approach does not use information from other years for registrations that are not operating; thus, log-linear models differ for each year as the number of registrations differs for each year. Furthermore, estimation with log-linear model tends to be more variable, especially for complete years.

6. Conclusion

In the present paper, a multidimensional Rasch model is proposed for the analysis of capture–recapture data. We assumed that registrations may be divided into two or more subgroups (not necessarily disjoint) measuring the latent variables accounting for correlations among registrations. As a consequence, the random variables denoting the presence or absence of an individual in a registration are assumed to be conditionally independent, given the latent variables.

Under the assumption that the posterior distribution of the latent variables follows a multivariate normal distribution, we applied the extension of the Dutch Identity proposed by Hessen [12] in a psychometric context to the capture–recapture framework, and we showed how to re-express the probability of a generic capture profile in terms of the log-linear multidimensional Rasch model. Then, we presented a re-parameterization of the proposed model that allows for a connection between the multidimensional Rasch model and the standard log-linear model. Applying these formulas, it is possible to compute the parameters of the standard log-linear model, starting from those of the multidimensional Rasch model. We also discussed an extension of the model for the situation in which a stratifying variable is available and the assumption of measurement invariance across strata can be made.

An application of the models discussed to the NTDs' data revealed that the final model for inference was one of the proposed log-linear multidimensional Rasch models. The final model was preferred over other log-linear models because it showed the smallest AIC and BIC values.

References

1. Seber GAF. *The Estimation of Animal abundance and Related Parameters*. Griffin: London, 1982.
2. International Working Group for Disease Monitoring and Forecasting, (IWGDMF). Capture–recapture and multiple-record systems estimation I: history and theoretical development. *American Journal of Epidemiology* 1995; **142**(10):1059–1068.
3. Fienberg SE. The multiple-recapture census for closed populations and the 2^k incomplete contingency tables. *Biometrika* 1972; **59**(3):591–603.
4. Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press: Cambridge, 1975.
5. Agresti A. Simple capture–recapture models permitting unequal catchability and variable sampling effort. *Biometrics* 1994; **50**(2):494–500.
6. Darroch N, Fienberg SE, Glonek GFV, Junker BW. A three-sample multiple-capture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* 1993; **88**(423):1137–1148.
7. Christensen KB, von Davier M, Lee Y. Testing unidimensionality in polytomous Rasch models. *Psychometrika* 2002; **67**(4):563–574.
8. Bartolucci F. A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika* 2007; **72**(2):141–157.
9. Haberman SJ, Davier M, Lee Y. Comparison of multidimensional item response models: multivariate normal ability distributions versus multivariate polytomous ability distributions. *ETS Research Report Series* 2008; **2**:i–25.
10. Bartolucci F, Forcina A. Analysis of capture–recapture data with a Rasch-type model allowing for conditional dependence and multidimensionality. *Biometrics* 2001; **57**:714–719.
11. Bartolucci F, Pennoni F. A class of latent Markov models for capture–recapture data allowing for time, heterogeneity, and behavior effects. *Biometrics* 2007; **63**(2):568–578.
12. Hessen DJ. Fitting and testing conditional multinormal partial credit models. *Psychometrika* 2012; **77**(4):693–709.
13. Holland PW. The Dutch Identity: a new tool for the study of item response model. *Psychometrika* 1990; **55**(1):5–18.
14. Reckase M. The difficulty of test items that measure more than one ability. *Applied Psychological Measurement* 1985; **9**:401–412.
15. Zwane EN, van der Pal K, van der Heijden PGM. The multiple- record systems estimator when registrations refer to different but overlapping populations. *Statistics in Medicine* 2004; **23**:2267–2281.
16. Zwane EN, Van der Heijden PGM. Analysing capture–recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations. *Statistics in Medicine* 2007; **26**:1069–1089.

Author Query Form

Dear Author,

During the copyediting of your paper, the following queries arose. Please respond to these by annotating your proof with the necessary changes/additions.

- If you intend to annotate your proof electronically, please refer to the E-annotation guidelines.
- If you intend to annotate your proof by means of hard-copy mark-up, please use the standard proofreading marks in annotating corrections. If manually writing corrections on your proof and returning it by fax, do not write too close to the edge of the paper. Please remember that illegible mark-ups may delay publication.

Whether you opt for hard-copy or electronic annotation of your proof, we recommend that you provide additional clarification of answers to queries by entering your answers on the query sheet, in addition to the text mark-up.

Query No.	Query	Remark
Q1	AUTHOR: "$n_{i_1 \ldots i_S j}$ and $\pi_{i_1 \ldots i_S j}$" has been changed to "$n_{i_1, \ldots, i_S j}$ and $\pi_{i_1, \ldots, i_S j}$". Please check if correct.	

USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

Required software to e-Annotate PDFs: <u>Adobe Acrobat Professional</u> or <u>Adobe Reader</u> (version 7.0 or above). (Note that this document uses screenshots from <u>Adobe Reader X</u>)
The latest version of Acrobat Reader can be downloaded for free at: <u>http://get.adobe.com/uk/reader/</u>

Once you have Acrobat Reader open on your computer, click on the Comment tab at the right of the toolbar:
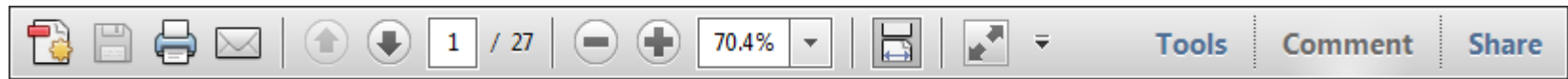
This will open up a panel down the right side of the document. The majority of tools you will use for annotating your proof will be in the Annotations section, pictured opposite. We've picked out some of these tools below:

---
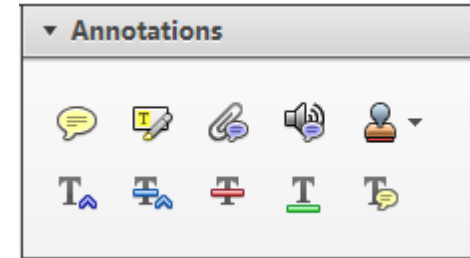
1. Replace (Ins) Tool – for replacing text.

Strikes a line through text and opens up a text box where replacement text can be entered.

How to use it

- Highlight a word or sentence.
- Click on the Replace (Ins) icon in the Annotations section.
- Type the replacement text into the blue box that appears.

---
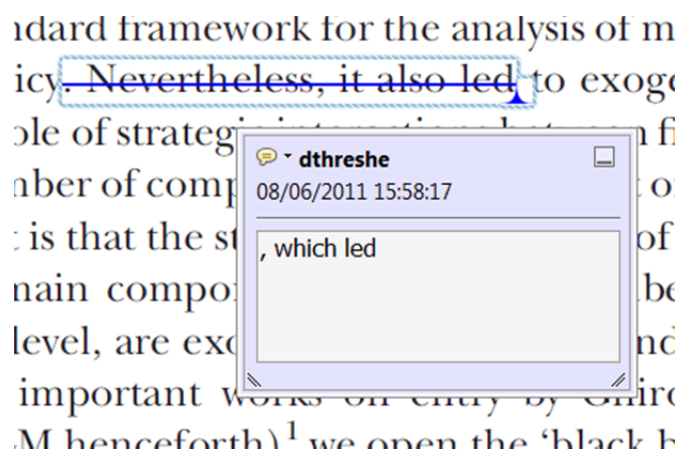
2. Strikethrough (Del) Tool – for deleting text.

Strikes a red line through text that is to be deleted.

How to use it

- Highlight a word or sentence.
- Click on the Strikethrough (Del) icon in the Annotations section.

---
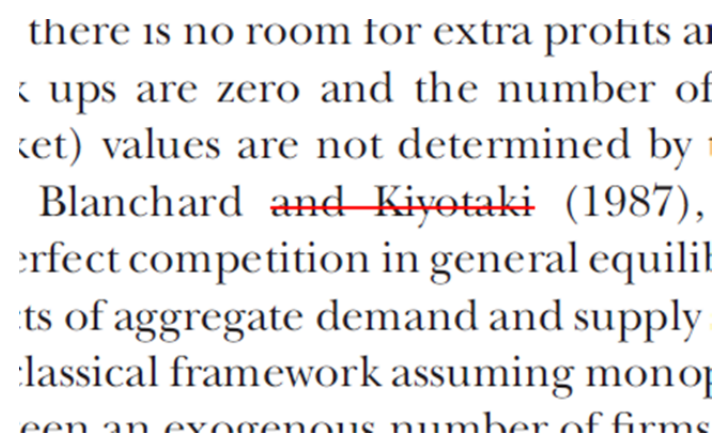
3. Add note to text Tool – for highlighting a section to be changed to bold or italic.

Highlights text in yellow and opens up a text box where comments can be entered.

How to use it

- Highlight the relevant section of text.
- Click on the Add note to text icon in the Annotations section.
- Type instruction on what should be changed regarding the text into the yellow box that appears.
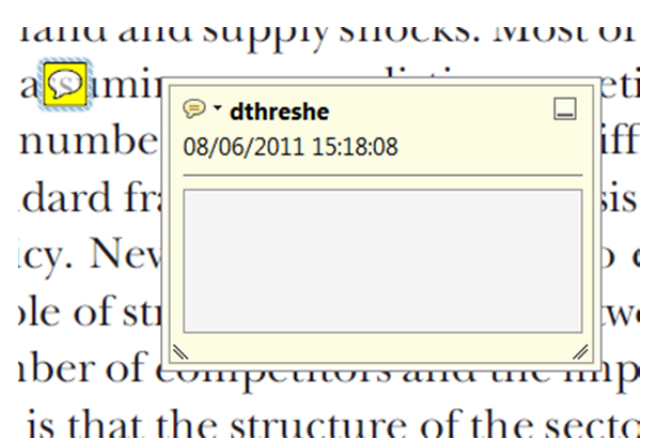
---
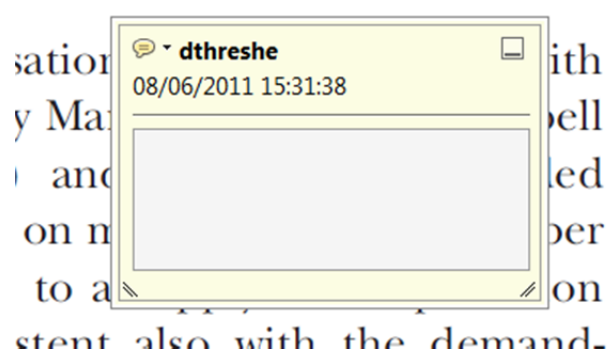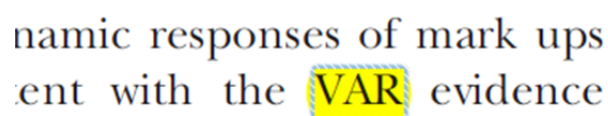
4. Add sticky note Tool – for making notes at specific points in the text.

Marks a point in the proof where a comment needs to be highlighted.

How to use it

- Click on the Add sticky note icon in the Annotations section.
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the yellow box that appears.

5. Attach File Tool – for inserting large amounts of text or replacement figures.

Inserts an icon linking to the attached file in the appropriate pace in the text.

How to use it

- Click on the Attach File icon in the Annotations section.
- Click on the proof to where you'd like the attached file to be linked.
- Select the file to be attached from your computer or network.
- Select the colour and type of icon that will appear in the proof. Click OK.
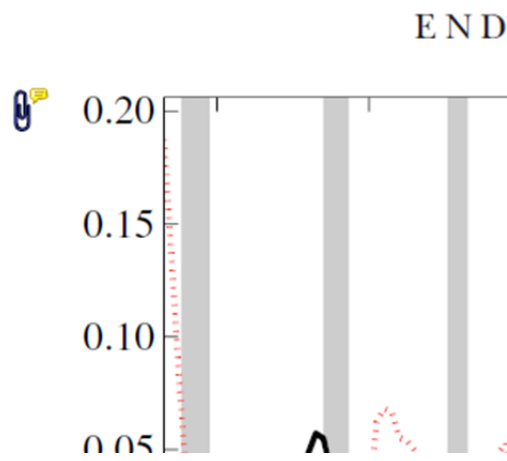
6. Add stamp Tool – for approving a proof if no corrections are required.

Inserts a selected stamp onto an appropriate place in the proof.

How to use it

- Click on the Add stamp icon in the Annotations section.
- Select the stamp you want to use. (The Approved stamp is usually available directly in the menu that appears).
- Click on the proof where you'd like the stamp to appear. (Where a proof is to be approved as it is, this would normally be on the first page).
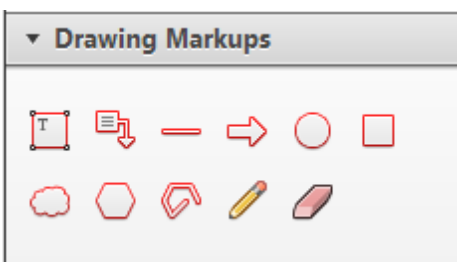
7. Drawing Markups Tools – for drawing shapes, lines and freeform annotations on proofs and commenting on these marks.

Allows shapes, lines and freeform annotations to be drawn on proofs and for comment to be made on these marks..

How to use it

- Click on one of the shapes in the Drawing Markups section.
- Click on the proof at the relevant point and draw the selected shape with the cursor.
- To add a comment to the drawn shape, move the cursor over the shape until an arrowhead appears.
- Double click on the shape and type any text in the red box that appears.
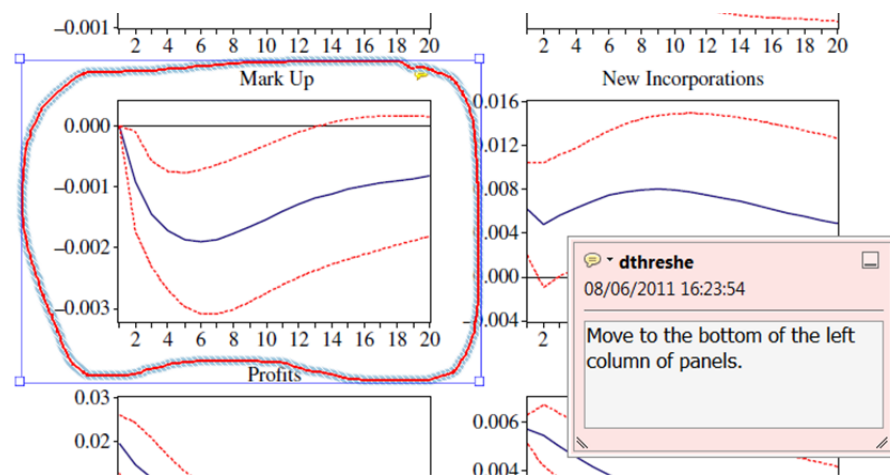
For further information on how to annotate proofs, click on the Help menu to reveal a list of further options: