Spanish Journal of Psychology (2014), **17**, e84, 1–8.
© Universidad Complutense de Madrid and Colegio Oficial de Psicólogos de Madrid
doi:10.1017/sjp.2014.81

A Derivation of the Polytomous Rasch Model Based on the Most Probable Distribution Method

Stefano Noventa[1], Luca Stefanutti[2] and Giulio Vidotto[2]

[1] *University of Verona (Italy)*
[2] *University of Padua (Italy)*

Abstract. Boltzmann's most probable distribution method is applied to derive the Polytomous Rasch model as the distribution accounting for the maximum number of possible outcomes in a test while introducing latent traits, item characteristics, and thresholds as constraints to the system. Affinities and similarities of the present result with other derivations of the model are discussed in light of the conceptual frameworks of statistical physics and of the principle of maximum entropy.

Received 21 June 2013; Revised 30 January 2014; Accepted 13 May 2014

Keywords: polytomous rasch model, principle of maximum entropy, most probable distribution.

Item Response Theory models (IRT, Lord & Novik, 1968) and Rasch Models (RM, Rasch, 1960) are fundamental in psychology. Their derivations generally assume monotonicity, continuity and asymptotical behavior of the item response function, plus general criteria or additional requirements like local stochastic independence, sufficiency of statistics, conditional inference or specific objectivity. The need for a dense set of items to account for interval scales has also been debated (Fischer, 1995a). For istance, the multidimensional Polytomous RM can be derived from sufficiency, while the Partial Credit model (PCM, Masters, 1982) and the Rating Scale model (RSM, Andrich, 1978; 1982) from conditional inference (Fischer, 1995b). Similarly, the conditional RM and the original RM can be derived requiring "measurement interchangeability" (Kelderman, 1995).

Derivations based on formal frameworks are relevant to gain insight on a model. The derivation based on conditional inference, for instance, describes both PCM and RSM as special cases of the Power Series Distribution. Similarly, the derivation based on items "interchangeability" (in respect to their relation to each other and with other variables) connects the RM to the Generalized Linear Models. Alternative derivations can also be given borrowing techniques from statistical physics: a derivation of the Polytomous RM based on the method of the steepest descent (Darwin & Fowler, 1922a; 1922b; 1923) was given by Ebneth (1993) considering a testee answering a fictional series of tasks. The average numbers of equivalent task series, conditional to constraints, were used to obtain the probability. A derivation of the dichotomous RM, based on Boltzmann's most probable distribution (MPD) method (see Huang, 1987) was instead given by Noventa, Stefanutti, & Vidotto (2013) conceiving a test as a distribution of responses with several possible outcomes constrained by means of latent traits and item characteristics. The RM was derived as the distribution maximizing the number of possible outcomes.

Although these two approaches to RM are based on different techniques, they are related by the principle of Maximum Entropy (MaxEnt). Indeed, they both converge to the asymptotic probability distribution that corresponds to the condition of maximum ignorance about the system (Jaynes, 1982).

In the present work, the derivation of the dichotomous RM given in Noventa et al. (2013) is extended to the Polytomous RM and its implications are discussed in light of the aforementioned literature, in particular, the Darwin-Fowler derivation (Ebneth, 1993), the MaxEnt principle and the use of a statistical physics framework for understanding the rationale of IRT and RM.

A brief introduction to the MPD method and the MaxEnt principle

The "most probable distribution" method

The MPD is a combinatorial method suggested by Boltzmann in 1877 to derive the energy distribution of particles in a gas. Particles are allocated into energy levels $\{\epsilon_k\}$ (like marbles in boxes) with occupation

Correspondence concerning this article should be addressed to Noventa Stefano. Center for Assessment. University of Verona. Verona (Italy).

E-mail: stefano.noventa@univr.it

The present work has been supported by a grant from the Center for Assessment, University of Verona, Director Prof. Giuseppe Favretto. We would also like to thank the anonymous reviewers of the journal for their keen insight into the manuscript and their unvaluable suggestions.

numbers $\{n_k\}$ under the constraints given by their total number N and energy E:

$$\sum_k n_k = N \qquad \sum_k n_k \epsilon_k = E \tag{1}$$

Since particles can be distributed in several ways, there are different possible combinations (and sets of frequencies $f_k = \frac{n_k}{N}$). Some of these combinations are also equivalent, since identical particles can be swapped. To understand which combination is more likely to be observed in Nature, *multiplicity* is defined as the number of ways in which a particular combination is realized:

$$W(\{n_k\}) = \frac{N!}{\prod_k n_k!} \tag{2}$$

The logarithm of the multiplicity is known as Boltzmann's Entropy. The gist of the MPD method is that the final probability distribution is associated to the combination that is realized in the maximum number of ways. To find such a probability the multiplicity (2), under the constraints (1), is maximed using the method of the Lagrangian multipliers:

$$\Lambda(\{n_k\}) = \ln W(\{n_k\}) + \lambda_1\left(N - \sum_k n_k\right) + \lambda_2\left(E - \sum_k n_k \epsilon_k\right) \tag{3}$$

were $\lambda_1, \lambda_2 \in \mathbb{R}$. The difficulty in maximizing (3) rely on the integer nature of $\{n_k\}$ and in the factorials contained in the multiplicity. An approximate derivation, based on the fact that the limit $N \to \infty$ is taken, requires Stirling's approximation for factorials and a continuous approximation for $\{n_k\}$. A better result, still in the continuity approximation, is obtained by replacing factorials with the Gamma function (Landsberg, 1954). An elegant derivation was given by Clinton and Massa (1972) avoiding both continuity and Stirling's approximations. Since the present work relies on this derivation, it is useful to briefly sketch it for the present case: Clinton and Massa noticed that $\Lambda(\{n_k\})$ is in a maximum if, moving a particle both in or out of an energy level, the following system of inequalities is satisfied:

$$\begin{cases} \Lambda(n_1, \dots, n_k, \dots) \geq \Lambda(n_1, \dots, n_k + 1, \dots) \\ \Lambda(n_1, \dots, n_k, \dots) \geq \Lambda(n_1, \dots, n_k - 1, \dots) \end{cases} \tag{4}$$

Substitution of (3) in the system (4) leads to:

$$\begin{cases} \ln(n_k + 1) + \lambda_1 + \lambda_2 \epsilon_k \geq 0 \\ \ln(n_k) + \lambda_1 + \lambda_2 \epsilon_k \leq 0 \end{cases}$$

so that any occupation number (once divided by N) is bounded by:

$$\frac{\exp(-\lambda_1)}{N}\exp(-\lambda_2 \epsilon_k) - \frac{1}{N} \leq \frac{n_k}{N} \leq \frac{\exp(-\lambda_1)}{N}\exp(-\lambda_2 \epsilon_k) \tag{5}$$

Since these are actually the occupation frequencies, for the squeeze theorem the probability can be defined taking the limit $N \to \infty$ where the terms $1/N$ vanishes:

$$p_k = \lim_{N\to\infty} f_k = \lim_{N\to\infty} \frac{\exp(-\lambda_1)}{N}\exp(-\lambda_2 \epsilon_k) \tag{6}$$

The condition $\sum_k \mathrm{p}_k = 1$ is used to introduce a normalization term:

$$\frac{N}{\exp(-\lambda_1)} = \sum_{\mathrm{k}} \exp(-\lambda_2 \epsilon_k) \tag{7}$$

and finally the probability is:

$$p_k = \frac{\exp(-\lambda_2 \epsilon_k)}{\sum_{\mathrm{k}} \exp(-\lambda_2 \epsilon_k)} \tag{8}$$

It is immediate the similarity of (8) with the Polytomous RM if the occupation numbers n_k are a measure of how many times a category response is chosen in N trials by a testee. Although with some differences in assumptions and methods, this line of reasoning is close to Ebneth's (1993). His derivation, however, requires the concept of *statistical ensemble*, a collection of N copies of the system that, independently and randomly, assume all the possible states allowed to the system (Huang, 1987). For instance, N copies of the *entire* gas. This concept is fundamental in statistical physics which is based on *averages* over ensembles, but it is not required in the MPD method (Landsberg, 1954).

Boltzmann's "molecular" perspective (the ensemble is the set of N particles) is followed in the present work, so that all the possible response patterns of the joint population of subjects and items are accounted (see section 3). Probability (8) is then derived as the most likely response pattern that can be observed in the population. The concept of MPD is well justified in a MaxEnt framework.

The principle of Maximum Entropy

Information Entropy is a measure of uncertainty (Shannon, 1948). The higher the entropy of a system the more unpredictable is its state. The concept was inspired by Gibbs statistical Entropy:

$$S_I = -\sum_k p_k \log p_k \tag{9}$$

and takes maximum value when the distribution is uniform. Connections between statistical entropy and information theory were higlighted by Jaynes (1957). In particular, he suggested the importance of a MaxEnt principle, briefly stated as "when we make inferences based on incomplete information we should drawn them from that probability distribution that has the maximum entropy permitted by the information we do have" (Jaynes, 1982). Such a method is a generalization of the usual methods of statistical inference that allows the choice of different priors. For instance, the MaxEnt solution for a set of probabilities $\sum_k p_k = 1$ over some events x_k is the uniform distribution (Laplace's principle of indifference). Adding some functions $f_j(x_k)$, whose expected values are constrained $E\left[f_j(x_k)\right] = a_j \in \mathbb{R}$, the result is the is the exponential family:

$$\max\left\{S_I \middle| f_j\right\} \to p_k = \frac{\exp\left(-\sum_j \lambda_j f_j(x_k)\right)}{\sum_m \exp\left(-\sum_j \lambda_j f_j(x_m)\right)} \qquad (10)$$

where $\lambda_j \in \mathbb{R}$. On the one side, the MaxEnt distribution works as the inverse of the Darmois-Koopman-Pitman theorem and creates a model for which the data are sufficient statistics (Jaynes, 1982). On the other side, it relates Gibbs and Boltzmann Entropies:

$$\lim_{N\to\infty} \frac{1}{N} \log W \approx \lim_{N\to\infty} \sum_k \frac{n_k}{N} \log \frac{n_k}{N} = -\sum_k p_k \log p_k = S_I \quad (11)$$

The probability distribution that maximizes entropy is numerically identical to the frequency that possesses the greatest multiplicity (Jaynes, 1982). Boltzmann's Entropy is the limit of Gibbs' Entropy if probabilities are equal. Working on the concept of *ensemble* Gibbs formulation is however more general and allows to describe interacting particles (Jaynes, 1965). Hence, the Boltzmann entropy seems to be a suitable description of RM since it requires independence between subjects and items.

Basic assumptions, definitions and notations

Measurement scale and equivalence classes

Given a set of subjects $\nu \in \{1, \ldots, s\}$, a set of items $i \in \{1, \ldots, m\}$, and a random variable, whose realizations are response categories $X_{\nu i} = x_{\nu i} \in \{0, \ldots, c\}$, the response matrix of a test is $\{x_{\nu i}\}$. In the dichotomous case, $c = 1$. A generalization of the response matrix to the population can be either a finite or an infinite matrix $\{x_{\nu i}\}$ with $\nu \in S$ and $i \in I$, where S and I are the populations of all the subjects and items. The union $P = S \cup I$, endowed with a weak order relation $\precsim$, allows comparisons between subjects and items. A common scale for latent traits and item parameters is a triple $\langle P, M, \phi \rangle$ where $\phi: \langle P, \precsim \rangle \to \langle M, \leq \rangle$, with $M \subseteq \mathbb{R}$, is an homomorphism that preserves the weak order (see for instance Krantz, Luce, Suppes, & Tversky, 1971; Luce, Krantz, Suppes, & Tversky, 1990; Suppes & Zinnes, 1963). Equivalence classes of all the subjects and of all the items possessing the same position on the measurement scale can be defined as:

$$S_\alpha = \left\{\nu \in S \colon \phi(\nu) = \alpha \in A \subseteq M\right\},$$
$$I_\delta = \left\{i \in I \colon \phi(i) = \delta \in D \subseteq M\right\} \qquad (12)$$

where $\alpha \in A$ and $\delta \in D$ are the values of the latent trait and of the item characteristic. Let *j* and *k* be the indexes spanning these sets, assuming they have at least countable cardinalities.

In order to build a Polytomous RM, thresholds are also needed. They are generally conceived as locations on the latent trait set that indicate a subject has exceeded a particular category response. Let $r \in \{0, \ldots, \mathrm{c}\}$ be the index spanning the category responses, hence τ_{kr} is the threshold value needed for scoring the category $x_{jk} = r$ in an item of characteristic δ_k. Thresholds appear then to be both levels of latent trait and item characteristic, $\tau_{kr} \in \mathrm{A} \cap \mathrm{D}$. In what follows, the situation is considered in which $\mathrm{A} = \mathrm{D}$, so that there is always a match between levels of latent trait and of item characteristic.

Probability

An important debate in IRT concerns the source of randomness. In the *stochastic subject view*, probability explains variations due to the person or to the test situation, in the *random sampling view*, probability is the proportion of subjects with the same latent trait giving positive answers (Moleenar, 1995). Another important debate concerns whether latent traits and item characteristics are on an ordinal level or on a metric continuum (Michell, 1990). In the former case, different subjects and items might be associated to the same non-metric scale value, in the latter, the probability of subjects and items to possess the same values would be zero. Interestingly, the MPD method accommodates all the previous perspectives. Equivalence classes (12) are defined independently on whether they describe an ordinal ranking or a coarse-grained description of a continuum (i.e, classes due to limited precision in measurement). Indeed, the MPD method moves from occupation numbers, so it is not important whether they result from resampling a subject or from different testees. In the most general perspective the response might be rewritten as $x_{jkr}^{\nu it}$ where the supra-indexes refer to subjects, items, and repetitions, while sub-indexes refer to latent traits, item characteristics, and category responses. Since debating about the existence of subjects with the same latent trait and item with the same characteristic is pointless in the MPD framework, ν, i, t will be dropped.

Let X_{jk} be the response variable corresponding to a subject with latent trait α_j, answering to an item with

https://doi.org/10.1017/sjp.2014.81 Published online by Cambridge University Press

characteristic δ_k, and let $x_{jk} = r \in \{0,\ldots,c\}$ be its realization. The matrix $\{x_{jk}\}$ can be ordered by increasing levels of latent trait and item characteristic and then partitioned into different blocks characterized by a couple (α_j, δ_k). The number of responses within each block is given by a set of numbers $\{n_{jkr}\}$. If $N_{jk} = \sum_{r=0}^{c} n_{jkr}$ is the total number of cells in the specific *jk*-th block, then the ratio n_{jkr}/N_{jk} gives the proportion of responses for the *r*-th category (the number of subjects giving a response *r* to a certain class of items or the proportion of responses of a single subject to a single item depending on the *view*). The probability of drawing from the population a testee with latent trait α_j, answering *r* (whose threshold is τ_{kr}) to an item with a characteristic value of δ_k, is then:

$$P\left(X_{jk} = r \middle| \alpha_j, \delta_k, \tau_{kr}\right) := \lim_{N_{jk}\to\infty} \frac{n_{jkr}}{N_{jk}} \tag{13}$$

where the limit accounts for an infinite population. The law of total probability becomes:

$$\sum_{r=0}^{c} P\left(X_{jk} = r \middle| \alpha_j, \delta_k, \tau_{kr}\right) = 1 \tag{14}$$

The most probable distribution for a Polytomous test

Permutations

Multiplicity can be derived consideing that the total number of ways in which n_{jkr} responses can fill the N_{jk} cells of the block is given by the binomial coefficient:

$$W_{jkr} = \binom{N_{jk}}{n_{jkr}} = \frac{N_{jk}!}{n_{jkr}!\left(N_{jk} - n_{jkr}\right)!} \tag{15}$$

Once the response category $r = 0$ has been filled, there is room left for $N_{jk} - n_{jk0}$ responses in the other categories, and so on. Hence:

$$W_{jk} = \binom{N_{jk}}{n_{jk0}} \times \binom{N_{jk} - n_{jk0}}{n_{jk1}} \times \ldots \times \binom{N_{jk} - \sum_{r=0}^{c-2} n_{jkr}}{n_{jkc-1}} \times 1$$

that, after some algebra, yields:

$$W_{jk} = \frac{N_{jk}!}{\prod_{r=0}^{c} n_{jkr}!} \tag{16}$$

so that all the possible ways in which all the blocks can be filled is given by:

$$W\left(\{n_{jkr}\}\right) = \prod_{jk} W_{jk} = \prod_{jkr} \frac{N_{jk}!}{n_{jkr}!} \tag{17}$$

that is exactly multiplicity (2) but generalized to the joint population of subjects and items.

Constraints

The first constraint that must be taken into account is given by the total number of cells in each block, that must sum up to the total number N of cells in the response matrix. It follows that:

$$N = \sum_{jk} N_{jk} = \sum_{jkr} n_{jkr} \tag{18}$$

The second constraint depends on the fact that the number of possible outcomes defined by the multiplicity (17) depends on α, δ, and τ. The number of responses n_{jkr} is conditional to latent traits, item characteristics and thresholds. The constraint can be modeled as an implicit function of α_j, δ_k, τ_{kr} and n_{jkr}. Namely, $H\left(\{n_{jkr}\}, \{\alpha_j\}, \{\delta_k\}, \{\tau_{kr}\}\right) = \mu$, with $\mu \in \mathbb{R}$. However, to avoid any interaction between different response categories in different blocks, additive independence is assumed:

$$H\left(\{n_{jkr}\}, \{\alpha_j\}, \{\delta_k\}, \{\tau_{kr}\}\right) = \sum_{jkr} h\left(n_{jkr}, \alpha_j, \delta_k, \{\tau_{kt}\}_{t\leq r}\right) = \mu \tag{19}$$

Notice that, in each block, constraints likely depend on all the thresholds $\{\tau_{kt}\}_{t\leq r}$ that precede the one associated to the *r*-th category. Any $h_{jkr} := h\left(n_{jkr}, \alpha_j, \delta_k, \{\tau_{kt}\}_{t\leq r}\right)$ is a generic constraint for the *r*-th category in the *jk*-th cluster, and is assumed to be a monotonic function of its arguments.

Derivation of the most probable distribution

The MPD can be derived by maximizing the Boltzmann Entropy given by (17) under the effect of constraints (18) and (19). As in section (2) this extremality problem under external constraints is reduced with Lagrangian multipliers method to the unconstrained maximization of the function:

$$\begin{aligned} \Lambda\left(\{n_{jkr}\}\right) = {} & \ln W\left(\{n_{jkr}\}\right) + \lambda_1\left(\sum_{jkr} n_{jkr} - N\right) \\ & + \lambda_2\left(\sum_{jkr} h\left(n_{jkr}, \alpha_j, \delta_k, \{\tau_{kt}\}_{t\leq r}\right) - \mu\right) \end{aligned} \tag{20}$$

where $\lambda_1, \lambda_2 \in \mathbb{R}$ and the sets of $\{\alpha_j\}$, $\{\delta_k\}$, $\{\tau_{kr}\}$ enter in the equation as parameters. As noticed, the variables n_{jkr} are positive integers so the previous equation is not differentiable. Since the derivation follows exactly the step given in section (2), apart from some minor changes, full proof is given in Appendix A. There is however a point worth to be mentioned: not all the shape of the constraint h_{jkr} do define a probability. A unique solution can be achieved when the constraints are linear functions of n_{jkr}, that is $h_{jkr} = n_{jkr} f_{jkr}$ where

$f_{jkr} := f(\alpha_j, \delta_k, \{\tau_{kt}\}_{t \le r})$ is a generic functions of latent traits, item characteristics and thresholds (addition of constants or functions unrelated to n_{jkr} does not affect the multiplicity, see Appendix A). Interestingly, linear constraints are related to averages of latent traits, thresholds and item characteristics over all blocks and category responses. They are indeed the expected values of the sufficient statistics for the exponential family as in equation (10).

The probability resulting from maximizing (20) is then:

$$P\left(X_{jk} = r \middle| \alpha_j, \delta_k, \tau_{kr}\right) = \frac{\exp\left(\lambda_2 f\left(\alpha_j, \delta_k, \{\tau_{kt}\}_{t \le r}\right)\right)}{\sum_{r=0}^{c} \exp\left(\lambda_2 f\left(\alpha_j, \delta_k, \{\tau_{kt}\}_{t \le r}\right)\right)} \quad (21)$$

since λ_1 becomes a normalization term λ_2 and is a scale factor. As it can also be easily noticed, PCM and RSM can be obtained from equation (21) setting the appropriate constraint (19) to:

$$H_{RSM}\left(\{n_{jkr}\}, \{\alpha_j\}, \{\delta_k\}, \{\tau_{kr}\}\right) = \frac{1}{\lambda_2} \sum_{jkr} n_{jkr} \left| \sum_{t=0}^{r} \left(\alpha_j - (\delta_k - \tau_{kt})\right) \right| \quad (22)$$

$$H_{PCM}\left(\{n_{jkr}\}, \{\alpha_j\}, \{\delta_k\}, \{\tau_{kr}\}\right) = \frac{1}{\lambda_2} \sum_{jkr} n_{jkr} \left| \sum_{t=0}^{r} \left(\alpha_j - \tau_{kt}\right) \right| \quad (23)$$

Similar constraints can be given for any form of the Polytmous RM (Rasch, 1961).

Discussion

In the present work Boltzmann's MPD method was used to obtain the Polytomous RM. This methodology was chosen for sake of simplicity: first, although methods like the steepest descent are preferable (Jaynes, 1982; Schrödinger, 1946), they lack the intuitivity of the MPD method. Darwin-Fowler's method, indeed, requires the concept of ensemble and knowledge of complex analysis. The MPD method instead needs few assumptions, is based on simple algebra and appears to be more suited for exploratory and introductory purpouses (Landsberg, 1954). Second, "the Boltzmann MPD method and the Darwin-Fowler method lead to the same result in the limit $N \to \infty$" (Jaynes, 1982; Schrödinger, 1946). They are equivalent in light of the MaxEnt principle, since the maximization of the total number of outcomes can be replaced by the maximization of the number of ways in which equally probable outcomes, conditional to the constraints, can be realized. Third, but not less important, it is argued that Boltzmann Entropy is a sufficient framework to describe the RM since it requires the independence of its basic elements, that is, subjects and items. The concept of ensemble, meaning a set of mental copies of the system assuming all its possible microstates, is however essential to extend the model to account for interactions. The concept however needs to be defined in the framework of psychology.

A possible solution was given by Ebneth (1993), in the *stochastic subject view*, considering a testee undergoing a series of fictional tasks as the ensemble. In the present work a different perspective is higlighted, that defines the concept of ensemble in relation to the idea of a test as a collection of responses of different subjects to several items. Following the ideas of section (3.2), a more general approach enclosing both *stochastic subject view* and *random sampling view* might be given considering an ensemble as the set of all the response matrices in the joint population of subjects and items.

From the perspective of statistical physics, different ensembles are possible: the *microcanonical* is the one in which states are equally likely since they possess the same energy; the *canonical* is the one in which some states are more likely than others since energy can vary; a *grand-canonical* is the one in which the total number of basic elements may also vary. Most importantly, the descriptions given by the different ensembles are equivalent in an infinite population, in which fluctuations are negligible, so that most of the microstates have the same average values of energy and of basic elements.

The equivalent of probability (21) can be derived as the probability associated to a canonical (C) or a grand-canonical (GC) statistical ensemble (Huang, 1987):

$$P_C(H_i) = \frac{\exp(-\beta H_i)}{\sum_i \exp(-\beta H_i)} \quad , \quad P_{GC}(H_i) = \frac{\exp\left(-\beta\left(\sum_j \mu_j N_{ij} - H_i\right)\right)}{\sum_i \exp\left(-\beta\left(\sum_j \mu_j N_{ij} - H_i\right)\right)} \quad (24)$$

In the case of a gas of particles, β is the Boltzmann factor associated to the temperature, H_i is the Hamiltonian (a function describing the energy) of the i-th microstate, N_{ij} is the number of particles of the j-th species in the i-th microstates and μ_{ij} is the chemical potential describing the energy related to exchanges of particles. In the case of a test a parallel might be to associate to β a discrimination factor, H_i to a function (the constraints) describing latent traits, item characteristics and thresholds, N_{ij} might be the number of subjects in the j-th category and μ_{ij} a related function describing changes in the number of subjects in a category (or in a formal similarity with the multidimensional polytomous RM they might be associated to person's value and weight parameter in the j-th latent trait).

Probabilities (24) are however conceptually different from (21). Boltzmann's MPD method is based on different assumptions, and probability (21) describes the

https://doi.org/10.1017/sjp.2014.81 Published online by Cambridge University Press

proportion of subjects in a response category (or the propensity distribution) for a given item difficulty. Probabilities (24) are instead descriptions of all the subjects in all the category responses of all the items. Hence, they do not provide distinct distributions for distinct blocks of the responses matrix, but the probability of an *entire* response matrix. The concept of ensemble allows then to introduce interactions whereas the Boltzmann approach requires independency: only if local stochastic independence holds the probabilites (24), by introducing an additive structure of the Hamiltonian like in constraint (19), can be decomposed into a product of probabilities (21). Similarly, the difference between a dichotomous and a polytomous RM is the absence of local independence between the category responses. Such a perspective generalizes RM to those situation in which local independence does not hold and interactions are required.

It is also important to notice that probability distributions (24) describe equilibrium solutions that satisfy the MaxEnt principle and account for the maximum uncertainty. As in the MPD method, they decribe a system in the microstate having the highest probability, that is a system whose entropy is in a maximum, or in other words, for which the knowledge of the observer is minimum. Exponential families are indeed MaxEnt solution in presence of linear constraints on the expected value of their sufficient statistics. For instance, probability (21) is the one that maximizes multiplicity (17) under linear constraints on expected values of latent traits, item characteristics and thresholds (see section 4.3). Interestingly, multiplicity (17) describes a series of independent categorical distributions (over r categories) each related to a block j, k, and within each block there are N_{jk} trials. The multinomial distribution (with fixed number of trials and unknown probabilities) in itself is an exponential family whose inverse parameter mapping is the generalization of the logistic function, called *softmax* function, that corresponds to equation (21). These concepts can also be traced back in other derivations: the fundamental one is based on sufficiency (Fischer, 1995b). The derivation based on conditional inference (Fischer, 1995b) connects the RM to the Power Series Distribution, that is an exponential family that contains the binomial, the geometric, and the Poisson distributions (Patil, 1965). Even the derivation based on "measurement interchangeability" (Kelderman, 1995) might have a parallel on the commutative nature of constraint (19) that allows the possibility of switching items.

Such a line of reasoning suggests that MaxEnt principle provides a rationale behind IRT models and RM in the framework of psychology (notice also that RM, rather than pairwise comparisons or psychometric logistic curve depends on which population is considered). RM would be the most suitable description of a test under the constraints previously described. Other IRT models related to exponential families would be descriptions of systems with different combinatorial natures, that is, situations in which a different state of information is available. Notice however that MaxEnt works with noiseless data, the complete case needs a full Bayesian approach (Jaynes, 1982).

Some final remarks. First, this result does not hold for a finite population, in which there is not a unique definition of probabilty (21). In such a case, fluctuations are not negligible and distributions describing different states should be considered. Second, such an approach to IRT and RM appears to separate the rationale behind the models from the problem of their measurement type. The derivation was indeed given independently on the nature of the measurement scale, given by the triple $\langle P, M, \phi \rangle$, and whose type depends on the admissible transformation of the homomorphism ϕ that maps subjects and items into latent traits and item characteristics as in the definition of equivalence classes (12). The MaxEnt principle indeed appears to justify the rationale behind the models, but does not grant a type of measurement that should be ascertained considering the permissible transformations allowed to the homorphism ϕ, by a specific constraint. This approach was applied in Noventa et al. (2013) to derive the measurement scale for the dichotmous RM under the requirement of a constraint satisfying specific objectivity. As a result, the metric degraded from an interval scale to an ordinal one depending on the cardinality of the equivalence classes (12). Finite equivalence classes granted indeed only an ordinal type, unless the constraint satisfied the axioms of conjoint measurement. Interestingly, this finding parallells the results of nonparametric item response models (see for instance, Karabatsos, 2001) in which cancellation axioms, up to the last empirically testable finite order, are required to obtain ordered-metric scales for respondents and items.The higher the cardinality of the sets of items, subjects, and category responses, the more the scales approximate an intervale scale.

In conclusion, the method of the MPD is a pratical and combinatorial way to derive the RM moving from assumptions of independence. The concept of ensemble and a statistical physics approach are however needed to generalize the system out of independence. Most of all, the MaxEnt principle suggests that RM and IRT models, might possess a deeper rationale in entropy. They would describe the distribution of responses that is more likely to find during an experiment.

https://doi.org/10.1017/sjp.2014.81 Published online by Cambridge University Press

References

Andrich D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. http://dx.doi.org/10.1007/BF02293814

Andrich D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, *47*, 105–113. http://dx.doi.org/10.1007/BF02293856

Aczel J. (1966). *Lectures on functional equations and their applications*. New York, NY: Academic Press.

Boltzmann L. (1877). Über die Beziehung dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht [On the relationship between the second main theorem of mechanical heat theory and the probability rates of the heat equilibrium].*Wiener Berichte*, *76*, 373–435.

Clinton W. L., & Massa L. J. (1972). Derivation of a statistical mechanical distribution function by a method of inequalities. *American Journal of Physics*, *40*, 608–610. http://dx.doi.org/10.1119/1.1988059

Darwin C. G., & Fowler R. H. (1922a). On the partition of energy. *Philosophical Magazine*, *44*, 450–479.

Darwin C. G., & Fowler R. H. (1922b). On the partition of energy – Part II Statistical principles and termodynamics. *Philosophical Magazine*, *44*, 823–842. http://dx.doi.org/10.1080/14786441208562558

Darwin C. G., & Fowler R. H. (1923). Fluctuations in an assembly in statistical equilibrium. *Proceedings of the Cambridge Philosophical Society*, *21*, 391–404.

Ebneth G. (1993). *Das Bruchzahlverständnis von Schülern: Eine Untersuchung mittels logistischer Modellbildung* [Students' understanding of fractions: an investigation using the logistic modeling]. Münster/New York, NY: Waxmann Verlag.

Fischer G. H. (1995a). Derivations of the Rasch Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models* (pp. 15–38). New York, NY: Springer-Verlag.

Fischer G. H. (1995b). Derivations of the Polytomous Rasch Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models* (pp. 293–305). New York, NY: Springer-Verlag.

Huang K. (1987). *Statistical mechanics*. New York, NY: John Wiley & Sons.

Jaynes E. T. (1957). Information theory and statistical mechanics. *The Physical Review*, *106*, 620–630. http://dx.doi.org/10.1103/PhysRev.106.620

Jaynes E. T. (1965). Gibbs vs. Boltzmann Entropies. *American Journal of Physics*, *33*, 391–398. http://dx.doi.org/10.1119/1.1971557

Jaynes E. T. (1982). On the rationale of Maximum-Entropy methods. *Proceedings of the EEE*, *70*, 939–952. http://dx.doi.org/10.1109/PROC.1982.12425

Karabatsos G. (2001). The Rasch Model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, *2*, 389–423.

Krantz D. H., Luce R. D., Suppes P., & Tversky A. (1971). *Foundations of measurement, additive and polynomial representations*. (Vol. 1). San Diego, CA: Academic Press.

Kelderman H. (1995). The Polytomous Rasch Model within the class of generalized linear symmetry models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models* (pp. 307–323). New York, NY: Springer-Verlag.

Landsberg P. T. (1954). On most probable distributions. *Proceedings of the National Academy of Sciences*, *40*, 149–154. http://dx.doi.org/10.1073/pnas.40.3.149

Lord F. M., & Novik M. R. (1968). *Statistical theories of mental test scores*. London, UK: Addison-Wesley Publishing Company.

Luce R. D., Krantz D. H., Suppes S., & Tversky A. (1990). *Foundations of measurement, Vol. 3: Representation, axiomatization and invariance*. San Diego, CA: Academic Press.

Masters G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*, 149–174. http://dx.doi.org/10.1007/BF02296272

Michell J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, MI: Erlbaum.

Molenaar I. W. (1995). Some background for Item Response Theory and the Rasch Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models* (pp. 3–14). New York, NY: Springer-Verlag.

Noventa S., Stefanutti L., & Vidotto G. (2013). An analysis of Item Response Theory and Rasch Models based on the most probable distribution method. *Psychometrika*. http://dx.doi.org/10.1007/s11336-013-9348-y

Patil G. P (1965). On the multivariate generalized power series distributions and its application to the multinomial and negative multinomial. In G. P. Patil (Ed.), *Classical and Contagiuos Discrete Distributions*, (pp. 183–194). London, UK: Pergamon Press.

Rasch G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenaghen, Denmark: The Danish Institute of Educational Research.

Rasch G. (1961). On general laws and the meaning of measurement in psychology. Proceedings of the IV. *Berkeley simposium on mathematical statistics and probability*, Vol IV (pp. 321–333). Berkeley, CA: University of California Press.

Schrödinger E. (1946). *Statistical thermodynamics*. Cambridge, UK: Cambridge University Press.

Shannon C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 623–656. http://dx.doi.org/10.1002/j.1538-7305.1948.tb00917.x

Suppes P., & Zinnes J. L. (1963). Basic theory of measurement. In R. D. Luce, R. R. Bush & E. Galanter (Edd.), *Handbook of Mathematical Psychology* (Vol. 1). New York, NY: Wiley.

Appendix A - Derivation of equation (21)

Following the derivation of Clinton and Massa (1972), equation (20) yield two inequalities:

$$\Lambda\left(n_{1r}, \ldots, n_{jkr}, \lambda_1, \lambda_2\right) \geq \Lambda\left(n_{1r}, \ldots, n_{jkr} + 1, \lambda_1, \lambda_2\right) \quad \text{(A1)}$$

$$\Lambda\left(n_{1r}, \ldots, n_{jkr}, \lambda_1, \lambda_2\right) \geq \Lambda\left(n_{1r}, \ldots, n_{jkr} - 1, \lambda_1, \lambda_2\right) \quad \text{(A2)}$$

https://doi.org/10.1017/sjp.2014.81 Published online by Cambridge University Press

In particular, dropping the sums and the indices *jk* for simplicity of notation, and dropping also the costant terms μ, *N* since thay cancel out, they become:

$$\ln\left(\frac{N!}{n_r!}\right)+\lambda_1 n_r+\lambda_2 h(n_r,\ldots)\geq \ln\left(\frac{N!}{(n_r+1)!}\right)+\lambda_1(n_r+1)+\lambda_2 h(n_r+1,\ldots)$$

$$\ln\left(\frac{N!}{n_r!}\right)+\lambda_1 n_r+\lambda_2 h(n_r,\ldots)\geq \ln\left(\frac{N!}{(n_r-1)!}\right)+\lambda_1(n_r-1)+\lambda_2 h(n_r-1,\ldots$$

so that, expanding the logarithm and simplifying the common terms become:

$$\ln(n_r+1)\geq \lambda_1+\lambda_2[h(n_r+1,\ldots)-h(n_r,\ldots)]$$

$$-\ln n_r \geq -\lambda_1-\lambda_2[h(n_r,\ldots)-h(n_r-1,\ldots)]$$

Once defined the forward and backward finite difference equations:

$$\Delta h_r = h(n_r+1,\ldots)-h(n_r,\ldots) \tag{A3}$$

$$\nabla h_r = h(n_r,\ldots)-h(n_r-1,\ldots) \tag{A4}$$

the previous inequalities (divided by *N*) yield upper and lower bounds in the proportion of responses given by the subjects to the *r*-th category into the *jk*-th block:

$$\frac{\exp(\lambda_1+\lambda_2\Delta h_r)}{N}-\frac{1}{N}\leq\frac{n_r}{N}\leq\frac{\exp(\lambda_1+\lambda_2\nabla h_r)}{N}$$

In an infinite population, namely in the limit $N\to\infty$, becomes:

$$\frac{\exp(\lambda_1+\lambda_2\Delta h_r)}{N}\leq\frac{n_r}{N}\leq\frac{\exp(\lambda_1+\lambda_2\nabla h_r)}{N}$$

so that frequencies are bounded by exponentials with different arguments, in contrast to equation (6). A unique definition of probability is given only if the condition $\Delta h_r = \nabla h_r$ holds. This is a functional equation that can be solved as a second-order linear homogeneous recurrence relation with constant coefficients (Aczel, 1966). Solutions have the shape $h_r = n_r f_r + g_r$ with f_r and g_r generic functions of latent traits, thresholds and item characteristic values. However, the term g_r cancels out and can be set equal to zero. Squeeze theorem and definition of probability (15) then yield:

$$P\left(X_{jk}=r\middle|\alpha_j,\delta_k,\tau_{kr}\right)=\lim_{N\to\infty}\frac{\exp(\lambda_1+\lambda_2 f_r)}{N} \tag{A5}$$

only if the limit converges to a finite value. This can be obtained by normalizing through the law of total probability (14), indeed substitution of (A5) in (14) yields:

$$\frac{\mathrm{N}}{\exp(\lambda_1)}=\sum_{\mathrm{r}}\exp(\lambda_2 f_r) \tag{A6}$$

So that once inserted (A6) in equation (A5) it yields as result probability (21).