Reliability of researcher capacity estimates and count data dispersion: a comparison of Poisson, negative binomial, and Conway-Maxwell-Poisson models

Boris Forthmann¹ · Philipp Doeblér²

Received: 10 August 2020 / Accepted: 5 January 2021 / Published online: 6 February 2021
© The Author(s) 2021

Abstract

Item-response models from the psychometric literature have been proposed for the estimation of researcher capacity. Canonical items that can be incorporated in such models to reflect researcher performance are count data (e.g., number of publications, number of citations). Count data can be modeled by Rasch's Poisson counts model that assumes equidispersion (i.e., mean and variance must coincide). However, the mean can be larger as compared to the variance (i.e., underdispersion), or b) smaller as compared to the variance (i.e., overdispersion). Ignoring the presence of overdispersion (underdispersion) can cause standard errors to be liberal (conservative), when the Poisson model is used. Indeed, number of publications or number of citations are known to display overdispersion. Underdispersion, however, is far less acknowledged in the literature. In the current investigation the flexible Conway-Maxwell-Poisson count model is used to examine reliability estimates of capacity in relation to various dispersion patterns. It is shown, that reliability of capacity estimates of inventors drops from .84 (Poisson) to .68 (Conway-Maxwell-Poisson) or .69 (negative binomial). Moreover, with some items displaying overdispersion and some items displaying underdispersion, the dispersion pattern in a reanalysis of Mutz and Daniel's (2018b) researcher data was found to be more complex as compared to previous results. To conclude, a careful examination of competing models including the Conway-Maxwell-Poisson count model should be undertaken prior to any evaluation and interpretation of capacity reliability. Moreover, this work shows that count data psychometric models are well suited for decisions with a focus on top researchers, because conditional reliability estimates (i.e., reliability depending on the level of capacity) were highest for the best researchers.

Keywords Researcher capacity · Item response models · Rasch Poisson count model · Conway-Maxwell-Poisson count model · Dispersion · Reliability

Mathematics subject classification 62P25

✉ Boris Forthmann
boris.forthmann@wwwu.de

¹ Institute of Psychology in Education, University of Münster, Münster, Germany

² Department of Statistics, TU Dortmund University, Dortmund, Germany

JEL classification C18

Introduction

Count data are part of the daily bread and butter of scientometric researchers. Perhaps most eminently, number of publications and number of citations are examined. A default statistical model for count data relies on the Poisson distribution. For example, a Poisson model was used to describe scientific productions of mathematicians, physicists, and inventors (Huber 2000; Huber and Wagner-Döbler 2001a, 2001b). However, in particular citation count distributions are known to have a smaller mean as compared to the variance (Didegah and Thelwall 2013; Ketzler and Zimmermann 2013). This phenomenon is known as overdispersion and it violates equidispersion (i.e., mean and variance coincide) as an important assumption for valid statistical inference based on the Poisson model (Hilbe 2011). This issue has been acknowledged in numerous scientometric studies (Didegah and Thelwall 2013; Ketzler and Zimmermann 2013; Sun and Xia 2016) with some of them dating back to the early 1980s (Cohen 1981). Underdispersion, however, has been less an issue in the scientometric literature. This is illustrated by a simple search in GoogleScholar (<https://scholar.google.com/>) that reveals 69 hits for the search string “source:scientometrics overdispersion” and 4 hits for the search string “source:scientometrics underdispersion” (retrieved on August the 4th in 2020). This huge difference can most likely be explained because conservative inference based on overestimated standard errors (Faddy and Bosch 2001) is perceived to be less problematic as compared to liberal inference.

Moreover, Forthmann et al. (2019a) have recently shown that such dispersion issues affect also the measurement precision of individual capacity estimates in item-response theory (IRT) models. In a large simulation study, they demonstrated that reliability estimates of person capacity derived from the Rasch Poisson count model (RPCM; i.e., an IRT model relying on the Poisson distribution; Rasch 1960) can be huge overestimates for an overdispersed population model. Analogously, RPCM-based reliability was found to be underestimated when data were simulated according to an underdispersed population model. In addition, also more complex population models with fractions of the items being overdispersed and underdispersed, respectively, resulted in mis-estimated RPCM-based reliabilities. These observations are critically important for applications of IRT count data models to quantify research performance. For example, the RPCM has been proposed as an elegant approach to measure researcher capacity based on a variety of bibliometric indicators (Mutz and Daniel 2018b). However, analogous to the known problems related to statistical inference, reliability for researcher capacity estimates derived from the RPCM can be heavily overestimated in the presence of overdispersion. Mutz and Daniel (2018b) found an impressive reliability estimate of 0.96 which might be misestimated. Hence, the aim of the current study is to investigate potentially misestimated measurement precision of researcher capacities by applying the RPCM, IRT count models based on the negative binomial distribution, and the more flexible Conway-Maxwell-Poisson count model in a large dataset of inventors and a reanalysis of Mutz and Daniel’s dataset.

Estimating researcher capacity by item response theory models

The measurement of research performance is critically important for many practical purposes such as allotment of funding, candidate selection for academic jobs, interim evaluations of researchers after starting a new position, management of research strategies at the institutional level (Mutz and Daniel 2018b; Sahel 2011; Zhang et al. 2019). Most often, single indicators such as the h index (Hirsch 2005) are used to quantify performance at the level of individual researchers. But a mere deterministic use of bibliometric indicators has shortcomings that can be overcome: Probabilistic accounts of bibliometric measures allow, for example, the quantification of measurement precision (Glänzel and Moed 2013). In this vein, it should be noted that stochastic variants of bibliometric indicators such as the h index exist (Burrell 2007). The contrast between deterministic vs. probabilistic measurement accounts is paralleled in the psychometric literature. IRT models are probabilistic in nature (Lord 1980) and used to link properties of test items (i.e., item parameters such as difficulty) and capacity of individuals (i.e., ability to solve the items) with observed behavior (i.e., has an item be solved by a person or not). Indeed, an IRT model from the Rasch family was used by Alvarez and Pulgarín (1996) to scale journal impact based on citation and publication counts. However, this approach did not resonate well in the scientometric literature (Glänzel and Moed 2013).

More recently, Mutz and Daniel (2018b) revived the idea to use IRT scaling for the assessment of researcher capacity and highlighted five problems with commonly used approaches to assess researcher capacity [for more details see (Mutz and Daniel 2018b) and the references therein]: (a) indicators are observed and do not represent an abstract latent variable reflecting a researcher's competency instead of mere observed performance, (b) measurement error (or precision) is not appropriately taken into account, (c) the count data nature of most performance indicators is not appropriately taken into account, (d) multiple indicators can be reduced to rather few underlying latent dimensions (suggesting that looking at many indicators for evaluation purposes might be unnecessary), and (e) indicators are potentially not comparable between different scientific fields. Mutz and Daniel (2018b) suggested to use Doeblér et al.'s (2014) extension of the RPCM to overcome these problems.

From the Rasch Poisson counts model to the Conway-Maxwell-Poisson counts model: gaining modeling flexibility

In the psychological literature, the RPCM was used to scale many different cognitive abilities such as reading competency (Jansen 1995; Jansen and van Duijn 1992; Rasch 1960), intelligence (Ogasawara 1996), mental speed (Baghaei et al. 2019; Doeblér and Holling 2016; Holling et al. 2015), or divergent thinking (Forthmann et al. 2016, 2018). The model was further used for migraine attacks (Fischer 1987) or sports exercises such as sit-ups (Zhu and Safrit 1993). All these examples have in common with most bibliometric indicators that they provide count data as responses for each of the respective test items (e.g., the number of correctly read words within an allotted test time). In this work, the RPCM and the other IRT models will be introduced and used within a Generalized Linear Mixed Model (GLMM) framework that allows linking the expected value with a log-linear term including a researcher capacity parameter and an item easiness parameter (De Boeck et al.

2011). In addition, a variance function exists for each of the models, i.e. a variance depends on an (item-specific) dispersion parameter and the person- and item-specific conditional mean. For example, in the RPCM the expected value μ_{ji} of count y_{ji} for person j and item i is modeled as a multiplicative function of capacity parameter $\varepsilon_j > 0$ and item easiness parameter $\sigma_i > 0$

$$\mu_{ji} = \varepsilon_j \sigma_i \quad (1)$$

Then, the mean in the RPCM is in fact log-linear:

$$\mu_{ji} = \varepsilon_j \sigma_i = \mu(\theta_j, \beta_i) = \exp(\beta_i + \theta_j) \quad (2)$$

for the log-transformed capacity parameter $\theta_j = \ln(\varepsilon_j)$ and item easiness parameter $\beta_i = \ln(\sigma_i)$. The higher β_i , the easier it is to obtain a comparably large item score. The probability to observe y_{ji} is Poisson, so the probability mass function is

$$P(Y_{ji} = y_{ji} | \varepsilon_j, \sigma_i) = \frac{\mu_{ji}^{y_{ji}}}{y_{ji}!} \exp(-\mu_{ji}) \quad (3)$$

This distributional assumption implies equidispersion, i.e. $\text{Var}(Y_{ji}) = E(Y_{ji})$. The dispersion parameter ϕ_i equals 1 for all items, and the variance function is $V_{ji}(\mu_{ji}, \phi_i) = \mu_{ji} \phi_i = \mu_{ji}$ for the RPCM.

Mutz and Daniel (2018b) used Doeblér et al.'s (2014) extension of the RPCM in which an upper asymptote for performance is proposed, intended for speed of processing measures in which simple cognitive tasks have to be solved under time constraints. The Doeblér et al. (2014) model is coined the Item Characteristic Curve Poisson counts model (ICCPM), because it borrows the sigmoid shape of Item Characteristic Curves in binary IRT, to describe the conditional means. Because of the known dispersion issues with bibliometric indicators (Didegah and Thelwall 2013; Ketzler and Zimmermann 2013; Sun and Xia 2016), Mutz and Daniel (2018b) used a Poisson-Gamma mixture extension of Doeblér et al.'s speed model (i.e., a negative binomial model). They performed a thorough simulation study on the parameter recovery of this model and recommended to use it only with large sample sizes ($N = 400$ or higher). Hence, to reduce sample size requirements for the purpose of the current work, we use the comparably less complex negative binomial model (Hung 2012) in which the expected value is modeled following Eq. 2, but replacing Eq. 3 by a negative binomial assumption. The variance function for this model is $V_{ji}(\mu_{ji}, \phi_i) = \mu_{ji} + \mu_{ji}^2 / \phi_i = \mu_{ji} + \mu_{ji}^2 / \exp(\tau_i)$ which allows item-specific overdispersion, but not underdispersion. It should be further noted, that this variant of the negative binomial distribution has been coined NB2 in the literature because of the quadratic nature of its variance parameterization (Hilbe 2011). This model will be referred to as NB2 counts model (NB2CM) in this work. For completeness, we will also consider a NB1 counts model (NB1CM; i.e., a model with a linear variance parameterization; Hilbe 2011) with $\mu_{ji} = \exp(\beta_i + \theta_j)$ and $V_{ji}(\mu_{ji}, \phi_i) = \mu_{ji}(1 + \phi_i) = \mu_{ji}(1 + \exp(\tau_i))$. The RPCM results as a border case of the NB1CM: when τ_i approaches $-\infty$, $\exp(\tau_i)$ approaches 0 for all items and hence $V_{ji}(\mu_{ji}, \phi_i)$ goes to μ_{ji} .

The third and final model considered in this work is the Conway-Maxwell-Poisson counts model (CMPCM; Forthmann et al. 2019a). The CMPCM extends the RPCM to allow for both overdispersion and underdispersion at the level of items. Hence, with respect to dispersion modeling the CMPCM can be considered the most flexible approach examined in this work. Importantly, the CMPCM is based on Huang's (2017) mean parameterization of the Conway-Maxwell-Poisson distribution that leads to a log-linear model

for the conditional mean as in Eq. 2. The conditional variance function for the CMPCM, however, cannot be provided in a simple formula. However, as it is the case for the other count models above, the variance is a function of the mean and dispersion parameter (see Huang 2017). Importantly, this parameterization is different from other suggested regression models based on the Conway-Maxwell-Poisson distribution (e.g., Guikema and Goffelt 2008; Sellers and Shmueli 2010), and is a bona fide generalized linear model (GLM; Huang 2017). Finally, in all four models, RPCM, NB2CM, NB1CM, and CMPCM, it is assumed that the capacity variable θ follows a marginal normal distribution with mean zero and standard deviation σ_θ^2 . This distributional assumption for θ implies that the item easiness parameters β_i can be interpreted as expected value for an item on the log-scale when researchers have average capacity of zero on log-scale. In addition, dispersion parameters τ_i reflect error variance and local reliability for the models based on the CMP and negative binomial distributions (concrete interpretation depends on the respective distributions). All model parameters are estimated by means of a marginal maximum likelihood (MML) approach which is common in the IRT literature (e.g., De Boeck et al. 2011).

Forthmann et al. (2019a) found in a simulation study that item easiness parameters and standard deviation of the latent capacity parameters in a CMPCM were already consistently estimated with sample sizes as small as $N=100$. In addition, item-specific dispersion parameter estimates were found to be only slightly underestimated (but increasing the number of items seemed to prevent this bias). Hence, sample size requirements for this flexible IRT model are less demanding as compared to Mutz and Daniel's (2018b) complex NB2 extension of the ICCPCM (Doeblér et al. 2014). This renders the CMPCM a potentially useful candidate model that needs to be examined in relation to negative binomial models that are perhaps most often applied in the scientometric literature to account for dispersion issues (e.g., Didegah and Thelwall 2013; Ketzler and Zimmermann 2013).

Reliability of researcher capacity estimates

Reliability has been defined within the framework of classical test theory (e.g., Gulliksen 1950). The basic equation of classical test theory proposes that an observed test score X results as the sum of a true score T and an error term E (i.e., $X=T+E$). Reliability refers to the ratio of true score variance to observed score variance or alternatively one minus the ratio of error variance to observed score variance. Hence, reliability estimates quantify measurement precision of test scores and have an intuitive metric ranging from zero to one. In the tradition of IRT, however, measurement precision of ability estimates refers to usage of standard errors and confidence intervals conditional on the ability level, but summary indices of measurement precision that are based on the information available conditional on ability were developed also for IRT. In this vein, reliability has been coined reliability of person separation, marginal reliability, or empirical reliability as early as in the 1980s (Green et al. 1984; Wright and Masters 1982) and it is defined as the ratio of estimated true ability variance adjusted for measurement error (i.e., error variance as quantified by the average squared standard error across all ability-specific squared standard errors) and the uncorrected estimated true ability variance. Hence, reliability here displays to some degree a conceptual similarity with reliability as it is defined within classical test theory (but it should not be confused with it). For example, error variance is constant in classical test theory, but it varies as a function of persons within IRT which requires averaging across error variances to yield a reliability estimate (Wang 1999). Alternatively, it can

be understood as the squared correlation (again this is analogous to classical test theory) between estimated ability parameters and the true ability parameters (Brown and Croudace, 2015; Brown 2018). This quantification of reliability is easy to calculate for a variety of available estimators of ability parameters (Brown and Croudace 2015; Brown 2018) and provides an established intuitive metric.

Notably, the empirical reliability estimates are based on the estimate of the variance of the capacity distribution and the standard errors of the capacity estimates (more details are provided in the method section below). Hence, biases in these estimates would directly result in misestimated empirical reliability. In a previous simulation study that focused on the CMPCM (Forthmann et al. 2019a), we found accurate estimates of the ability variance and accurate standard errors even for sample sizes as small as $N=100$. Parameter recovery for the CMPCM has not been examined yet for sample sizes smaller than $N=100$, hence for situations in which only smaller samples are available new simulations should be run to examine potential bias of reliability estimates. Furthermore, if simulation findings generalize to the negative binomial models is also an open question and requires attention for situations with smaller sample sizes. As a more general remark, it should be noted that reliability is population specific and invariance across samples is not guaranteed (e.g., samples that are affected by range restriction may not result in an accurate estimate of capacity variance). Finally, measurement precision of reliability estimates can be assumed to be a function of the value of reliability with wider confidence intervals for reliability around the 0.50 s and quite narrow confidence intervals for excellent reliability above 0.90. Analytically, this is guaranteed by results of Feldt et al. (1987) and valid for cases when Cronbach's α and reliability coincide.

Aim of the current study

The goal of this study is a thorough comparison of various available count data IRT models (i.e., RPCM, NB2CM, NB1CM, and CMPCM) based on two scientometric datasets. All of these models allow modeling of the expected value in a log-linear fashion. The models differ with respect to their capability to model dispersion in the data. The RPCM is the least flexible model that is based on the assumption of equidispersion. The NB2CM and NB1CM allow global or item-specific overdispersion modeling, whereas the CMPCM is the most flexible model allowing global or item-specific overdispersion and/or underdispersion. Hence, the first goal is to compare these models in terms of their relative fit to the data based on information criteria that also take model parsimony into account. In addition, a series of likelihood ratio tests are used to compare nested models of increasing complexity. The second and main aim of this work is to examine the reliability of the researcher capacity estimates for the best fitting model and the other competing candidate models. Standard errors are known to be liberal (conservative) when data are overdispersed (underdispersed) which biases statistical inference based on the simple Poisson model. However, Forthmann et al. (2019a) demonstrated that these known problems further affect the reliability of capacity estimates. The RPCM was found to overestimate reliability with an overdispersed simulation model and an underdispersed population model resulted in underestimation of capacity reliability estimates. Hence, this work extends the promising work by Mutz and Daniel (2018b) in two important ways: a) comparing a greater variety of distributional models, b) examining less complex models with lower demands in terms of

sample size, and c) providing a detailed check of the inter-relatedness of capacity reliability estimates and whether or how dispersion was taken account.

Method

Data sources

Patent dataset

The first dataset is a subset of the patent dataset provided by the National Bureau of Economic Research (<https://data.nber.org/patents/>). The full dataset is described in Hall et al. (2001) and we use the file in which inventors were identified by an disambiguation algorithm (Li et al. 2014). The disambiguated data are openly available at the Harvard Dataverse (<https://dataverse.harvard.edu/dataverse/patent>). Here we use the same subset of $N=3055$ inventors that was used by Forthmann et al. (2019b). To control for issues arising from data truncation only inventors with careers within the years from 1980 to 2003 were used. Moreover, inventors were required to have at least one patent in each of the six four-year intervals in which the data were split. In this study, we use the number of patents granted as an indicator of inventive capacity. Hence, the number of patents for each of the six four-year intervals, respectively, was used as one item in this study (i.e., six items in total). It has been argued that productivity within a period of time (annual productivity is perhaps most often used in this regard) is highly relevant as a measure of scientific excellence (Yair and Goldstein 2020).

Mutz and Daniel's dataset

The second dataset comprises of $N=254$ German social sciences researchers who were listed in a membership directory of a quantitative methodology division of an unspecified academic society (Mutz and Daniel 2018b). The following six bibliometric indicators measure researcher capacity (Mutz and Daniel 2018b): a) TOTCIT: total number of citations received (excluding the highest cited paper), b) SHORTCIT: number of citations received within a 3-year citation window, c) NUMPUB: total number of published articles, d) TOP10%: number of publications in the top 10% of the researcher's scientific field, e) PUBINT: the number of papers published together with international co-authors, and f) NUMCIT: the number of papers that received at least one citation. The dataset is openly available for reanalysis (Mutz and Daniel 2018a).

Analytical approach

All models were fitted with the statistical software R (R Core Team 2019) by means of the glmmTMB package (Brooks et al. 2017). All R scripts and links for download of the datasets are provided in the online repository of this work (<https://osf.io/em642/>). All models were fitted with the same log-linear model to predict the expected value (see Eq. 2) and a normally distributed capacity parameter θ on the log-scale. The mean of the capacity parameter distribution was fixed to a value of zero and the variance σ_{θ}^2 was estimated (see Forthmann et al. 2019a). The glmmTMB package provides empirical Bayes estimates for

each θ_j by means of the maximum a posteriori (MAP) estimator. The NB2CM, NB1CM, and CMPCM were fit in two variants. First, more parsimonious models with only one dispersion parameter for all items were estimated. Then, models with item-specific dispersion parameters were fit. Dispersion parameters in glmmTMB are modeled with a log-link. For the negative binomial models this implies that large negative estimated values imply the absence of overdispersion. For the CMPCMs, however, negative values imply underdispersion, whereas a value of zero implies equidispersion and positive values imply

Table 1 Patent data: Model estimation results for RPCM and CMPCMs

	RPCM	NB2CM with global disper- sion	NB2CM with item-specific dispersion	CMPCM with global disper- sion	CMPCM with item-specific dispersion
Fixed effects	β (SE_β)	β (SE_β)	β (SE_β)	β (SE_β)	β (SE_β)
1980–1983	0.719 (0.015)***	0.769 (0.017)***	0.760 (0.016)***	0.762 (0.018)***	0.750 (0.015)***
1984–1987	1.192 (0.013)***	1.221 (0.016)***	1.218 (0.015)***	1.219 (0.016)***	1.220 (0.015)***
1988–1991	1.414 (0.013)***	1.429 (0.015)***	1.428 (0.015)***	1.434 (0.015)***	1.438 (0.015)***
1992–1995	1.504 (0.012)***	1.507 (0.015)***	1.511 (0.016)***	1.519 (0.015)***	1.525 (0.016)***
1996–1999	1.438 (0.013)***	1.455 (0.015)***	1.461 (0.016)***	1.458 (0.015)***	1.467 (0.016)***
2000–2003	1.029 (0.014)***	1.075 (0.016)***	1.081 (0.017)***	1.064 (0.017)***	1.072 (0.017)***
Random effects					
$\hat{\sigma}_\theta^2$	0.259	0.208	0.204	0.218	0.206
$\frac{\hat{\sigma}_\theta^2}{SE_\theta^2}$	0.042	0.064	0.063	0.067	0.066
Empirical reli- ability	0.837	0.694	0.689	0.692	0.682
Dispersion	τ	τ (SE_τ)	τ (SE_τ)	τ (SE_τ)	τ (SE_τ)
Global	0	1.411 (0.022)***		1.015 (0.020)***	
1980–1983			1.883 (0.089)***		0.243 (0.048)***
1984–1987			1.525 (0.062)***		0.805 (0.051)***
1988–1991			1.505 (0.058)***		0.960 (0.050)***
1992–1995			1.246 (0.049)***		1.400 (0.058)***
1996–1999			1.261 (0.050)***		1.353 (0.058)***
2000–2003			1.275 (0.055)***		1.123 (0.062)***
Model comparison					
$\Delta\chi^2(df)$	–	6717.07 (1)***	58.90 (5)***	4792.74 (1)***	231.66 (5)***
AIC	91,533.90	84,818.82	84,769.92	86,743.15	86,521.49
BIC	91,588.61	84,881.36	84,871.54	86,805.68	86,623.11
Akaike weights	0	0	1	0	0

$N=3055$ (18,330 observations); τ =dispersion parameter (CMPCMs: $\tau<0$ indicates underdispersion; $\tau=0$ indicates equidispersion; and $\tau>0$ indicates overdispersion; NB2CMs: zero is not a useful reference for τ). Likelihood ratio tests for NB1CMs: NB1CM with global dispersion is compared with RPCM and NB1CM with item-specific dispersion is compared with CMPCM with global dispersion. Likelihood ratio tests for CMPCMs: CMPCM with global dispersion is compared with RPCM and CMPCM with item-specific dispersion is compared with CMPCM with global dispersion. $\Delta\chi^2$ =likelihood ratio statistic; AIC=Akaike's Information Criterion; BIC=Bayesian Information Criterion. Lower values of information criteria imply better model fit when also model parsimony is taken into account. BIC values model parsimony more than AIC. Values in bold indicate the best fitting model based on AIC and BIC, respectively

* $p<.05$; ** $p<.01$; *** $p<.001$

Table 2 Data from Mutz and Daniel (2018a, b): Model estimation results for RPCM and CMPCMs

	RPCM	NB1CM with global disper- sion	NB1CM with item-specific dispersion ^a	CMPCM with global disper- sion	CMPCM with item-specific dispersion
Fixed effects	$\beta (SE_{\beta})$	$\beta (SE_{\beta})$	$\beta (SE_{\beta})$	$\beta (SE_{\beta})$	$\beta (SE_{\beta})$
TOTCIT	3.900 (0.130)***	4.178 (0.109)***	4.688 (0.108)***	4.415 (0.117)***	3.946 (0.124)***
SHORTCIT	3.079 (0.130)***	3.383 (0.111)***	3.868 (0.103)***	3.333 (0.119)***	3.518 (0.116)***
NUMPUB	1.542 (0.130)***	1.940 (0.117)***	2.330 (0.096)***	2.169 (0.124)***	2.194 (0.101)***
TOP10%	-0.542 (0.135)***	0.291 (0.134)*	0.247 (0.110)*	-0.201 (0.142)	0.076 (0.116)
PUBINT	0.376 (0.132)**	0.908 (0.126)***	1.164 (0.105)***	0.811 (0.132)***	0.995 (0.111)***
NUMCIT	1.298 (0.131)***	1.764 (0.118)***	2.085 (0.096)***	1.835 (0.125)***	1.951 (0.100)***
Random effects					
$\hat{\sigma}_{\theta}^2$	4.198	2.824	1.868	3.169	2.478
$\overline{SE}_{\theta}^2$	0.078	0.119	0.082	0.182	0.067
Empirical reli- ability	0.981	0.958	0.956	0.942	0.973
Dispersion	τ	$\tau (SE_{\tau})$	$\tau (SE_{\tau})$	$\tau (SE_{\tau})$	$\tau (SE_{\tau})$
Global	0	2.208 (0.060)***		3.494 (0.020)***	
TOTCIT			5.106 (0.106)***		8.566 (1.934)***
SHORTCIT			3.717 (0.099)***		5.260 (0.312)***
NUMPUB			-25.306 (0.129)***		-0.322 (0.116)**
TOP10%			0.214 (0.234)		1.260 (0.230)***
PUBINT			0.963 (0.161)***		1.695 (0.189)***
NUMCIT			-25.306 (0.129)***		-2.173 (0.169)***
Model comparison					
$\Delta\chi^2(df)$	—	6102.10 (1)***	1483.00 (5)***	6932.91 (1)***	700.34 (5)***
AIC	17,579.28	11,479.16	10,004.21	10,648.37	9958.03
BIC	17,616.58	11,521.80	10,068.16	10,691.00	10,027.31
Akaike weights	0	0	0	0	1

$N=254$ (1524 observations); τ =dispersion parameter (CMPCMs: $\tau<0$ indicates underdispersion; $\tau=0$ indicates equidispersion; and $\tau>0$ indicates overdispersion; NB2CMs: zero is not a useful reference for τ). Likelihood ratio tests for NB1CMs: NB1CM with global dispersion is compared with RPCM and NB1CM with item-specific dispersion is compared with CMPCM with global dispersion. Likelihood ratio tests for CMPCMs: CMPCM with global dispersion is compared with RPCM and CMPCM with item-specific dispersion is compared with CMPCM with global dispersion. $\Delta\chi^2$ =likelihood ratio statistic; AIC= Akaike's Information Criterion; BIC= Bayesian Information Criterion. Lower values of information criteria imply better model fit when also model parsimony is taken into account. BIC values model parsimony more than AIC. ^aThe item dispersion parameter estimates for Item 3 and Item 6 were constrained to be equal for the NB1PCM with item-specific dispersion

* $p<.05$; ** $p<.01$; *** $p<.001$

overdispersion (Forthmann et al. 2019a). The tables in which model results are reported include a note on these different interpretations of the dispersion parameters to facilitate interpretation (see Tables 1 and 2). Models with increasing complexity were compared based on likelihood ratio tests (denoted by $\Delta\chi^2$) and the Akaike information criterion (AIC; Akaike 1973) and the Bayesian information criterion (Schwarz 1978). The information criteria take also model parsimony into account with the BIC imposing a stronger penalty for complex models. We also calculated Akaike weights for multi-model inference as implemented in the R package MuMIn (Barton 2019). For the NB2PCMs and the NB1PCMs we first looked at model comparison statistics and—for the sake of brevity—report here only the respective better fitting variants of the negative binomial models. Results for the not reported models can be found in the online repository for this work (<https://osf.io/em642/>).

Reliability of the researcher capacity estimates was globally determined based on empirical reliability (Brown and Croudace 2015; Green et al. 1984):

$$Rel(\theta) = 1 - \overline{SE}_\theta^2 / \hat{\sigma}_\theta^2 \quad (4)$$

with $\overline{SE}_\theta$ being the average standard error of the researcher capacity estimates and $\hat{\sigma}_\theta^2$ the estimated variance of the researcher capacity distribution. In addition, conditional reliability (i.e., the reliability for a specific capacity level) can be calculated analogously to Eq. 4 (Green et al. 1984):

$$Rel(\theta_j) = 1 - SE_{\theta_j}^2 / \hat{\sigma}_\theta^2 \quad (5)$$

with SE_{θ_j} being the standard error of specific capacity estimate θ_j . We compared conditional reliability estimates to explore the dependence of reliability on the capacity level. Our main aim in this regard was to compare both empirical and conditional reliability estimates between the respective best fitting model and the alternative models to reveal how model selection might influence the evaluation of reliability and important related interpretations (e.g., deciding that estimates are accurate enough for a high-stakes decision).

Results

Patent dataset

The parameter estimates of all models can be found in Table 1 [with the exception of the NB1PCMs (global dispersion: AIC=86,680.06, BIC=86,742.59; item-specific dispersion: AIC=86,247.58, BIC=86,349.19) that were found to fit less well to the data as compared to the NB2PCMs (global dispersion: AIC=84,818.82, BIC=84,881.36; item-specific dispersion: AIC=84,769.92, BIC=84,871.54)]. The item easiness parameter estimates were highly comparable across all fitted models. These parameters indicate an increase of productivity up to the interval from 1992 to 1995. Then, productivity seemed to decrease slightly up to the final intervals of inventors' careers, but productivity did not fall back to the level of the first career interval. All models that take deviations from the Poisson assumption of equidispersion into account fitted better than the RPCM (as indicated by likelihood ratio tests and information criteria; see Table 1). All models (i.e., those with general and item-specific dispersion parameters) clearly indicated the presence of overdispersion.

The overall best fitting model was the NB2PCM with item-specific dispersion parameters. The estimated dispersion parameters indicated that overdispersion was strongest for the first career interval. In addition, the amount of overdispersion was found to decrease monotonically over inventor's careers (see Table 1). It is noteworthy that this pattern of dispersion parameters was not found to be paralleled by the CMPCM with item-specific dispersion. To examine whether this observation was masked by the complex interplay between conditional mean and dispersion, we checked the item-specific dispersion index [$Var(Y_i)/E(Y_i)$; Bonat et al. 2018; Consul and Famoye 1992] with $\theta_j=0$ (i.e., average capacity). Sellers and Shmueli's (2010) approximate variance formula was used for the calculation of item-specific dispersion indices (which was justified because all ν were <1). However, dispersion indices did also yield a different pattern as compared to the NB2CM item-specific dispersion parameters. That is, the findings revealed that the dispersion index increased from 1.18 to 2.58 at the fourth time interval (1992–1995) and, then, decreased to a value of 1.90 for the last interval (2000–2003).

Furthermore, as expected when overdispersion is present, the RPCM led to a huge overestimation of empirical reliability of capacity estimates (0.837) as compared to any of the other models that take overdispersion into account (range of empirical reliability estimates: 0.682–0.694). This overestimation resulted from both an overestimation of the variance of the capacity estimate distribution as well as an underestimation of the average standard errors of the capacity estimates (see Table 2). While empirical reliability estimates were found to be highly comparable across all models with dispersion parameters (see Table 2), this was not the case for conditional reliability (see Fig. 1). In Fig. 1, conditional reliability estimates are plotted against the z -transformed capacity estimates for all count models. Conditional reliability based on the RPCM was clearly overestimated for the full range

Fig. 1 Bivariate scatterplot of the conditional reliability estimates for the respective researcher capacity estimates (y axis) against the z -standardized capacity estimates (x axis) for the patent data

of capacity estimates. In addition, conditional reliability estimates for the NB2PCMs and CMPCMs were quite comparable up to a capacity of 1.5 *SDs* above the mean. For values greater than 1.5 *SDs* above the mean, it is clearly visible in Fig. 1 that the CMPCMs overestimate conditional reliability for highly productive inventors (this interpretation is only feasible in comparison to the NB2PCM with item-specific dispersion as the best fitting model for this dataset). The NB2CM with item-specific dispersion displayed also some slightly lower conditional reliabilities in the top capacity range as compared to the NB2CM with global dispersion parameter. Hence, particularly for the most productive inventors in this sample, the choice of the distributional model and approach to dispersion modeling was crucial for an accurate assessment of reliability of capacity estimates.

Mutz and Daniel's (2018a, b) dataset

All parameter estimates of all fitted models (with the exception of the NB2PCM) can be found in Table 2. The NB2PCM with global dispersion parameter ($AIC=10,350.75$; $BIC=10,393.38$) fitted better as compared to the NB1PCM with global dispersion parameter ($AIC=11,479.16$; $BIC=11,521.80$). However, for this dataset we chose the NB1PCMs above the NB2PCMs because the NB1PCM with item-specific dispersion ($AIC=10,004.21$; $BIC=10,068.16$) fitted better to the data as compared to the NB2PCM with item-specific dispersion (the model estimation had many technical problems so that information criteria could not be calculated) and, thus, provided stronger competition for the overall model comparison procedure. In addition, we had to fix the item-specific dispersion parameters for NUMPUB and NUMCIT to the same value to deal with generally observed technical problems with the negative binomial models with item-specific dispersion. Notably, these problems were not unexpected because Mutz and Daniel (2018b) found that the dispersion parameters for these items adhered to the Poisson model.

The absolute values of the item-easiness parameter estimates (see Table 2) across the fitted models differed considerably stronger as compared to the patent dataset (see Table 1). Both models with a general dispersion parameter displayed overdispersion and fitted better as compared to the RPCM (see Table 2). The order of the dispersion parameter estimates in both models with item-specific dispersion parameters was highly comparable. However, the CMPCM with item-specific dispersion displayed underdispersion for NUMPUB and NUMCIT, whereas the large negative values for log-dispersion in the NB1PCM for these two items indicated that dispersion was at the lower limit of the parameter space (i.e., these items adhered to the Poisson model; see also Mutz and Daniel 2018b). The CMPCM with item-specific dispersion with unambiguously the best fitting model across all criteria. This observation is crucial because it highlights the inability of negative binomial models to take underdispersion into account. Clearly, the presence of underdispersion here caused technical problems for the estimation of negative binomial models. In addition, it is important to consider that IRT models deal with conditional distributions that can display underdispersion even when the unconditional distribution of, for example, the number of publications does not display underdispersion.

Empirical reliability estimates for this dataset were much more comparable across fitted models (range from 0.942 to 0.981; see Table 2) as compared to the patent data (see Table 1). Again, the highest reliability estimate resulted for the RPCM. However, the best fitting model here produced a reliability estimate of almost the same size (i.e., 0.973). Hence, even in situations in which the data require item-specific dispersion

modeling, it can be the case that compared to the RPCM reliability appears to be pretty accurately estimated because some items may display overdispersion (here: TOTCIT, SHORTCIT, TOP10%, and PUBINT) and some items may display underdispersion (here: NUMPUB and NUMCIT). Nonetheless, the variance of the capacity distribution was clearly overestimated in the RPCM as compared to the CMPCM with item-specific dispersion and also the average standard error was slightly overestimated. All other models resulted in higher average standard errors because these models all modeled overdispersion (i.e., the negative binomial models can only model overdispersion and the CMPCM with global dispersion empirically demonstrated overdispersion).

Figure 2 shows the conditional reliability plot for Mutz and Daniel’s dataset. The plot clearly shows that both NB1CMs and the CMPCM with global dispersion parameter would have led to strong underestimates of conditional reliability across almost the full range of capacity. In addition, the differences between the best fitting CMPCM with item-specific dispersion and all other dispersion models decreases with increasing capacity. The RPCM as compared to the best fitting CMPCM with item-specific dispersion tends to overestimate conditional reliability slightly stronger towards the lower tail of capacity (see Fig. 2).

Fig. 2 Bivariate scatterplot of the conditional reliability estimates for the respective researcher capacity estimates (y axis) against the z-standardized capacity estimates (x axis) for the BQ data (Mutz and Daniel 2018b)

Discussion

Accurate evaluations of researcher capacity are of high practical value in several contexts requiring selection decisions (e.g., for academic jobs, funding allotment, or research awards). The attractiveness of IRT models in this context has been highlighted in recent research (Mutz and Daniel 2018b) and the current work extends this idea in important ways. First, less complex models to predict the expected values were considered to reduce requirements with respect to sample size. Second, a model based on the Conway-Maxwell-Poisson distribution was added to the pool of candidate distributions. This distribution is more flexible as compared to the frequently used negative binomial distribution as it can handle equidispersion, overdispersion, and underdispersion at the item level. Finally, this study focused on the reliability of researcher capacity and the complex interplay of the chosen distributional model and conditional reliability (i.e., measurement precision at specific capacity levels.).

This study further amplifies the call for item-specific dispersion modeling because across the studied datasets models with item-specific dispersion fitted best. Moreover, it has been demonstrated that CMPCMs are a useful alternative model for bibliometric indicators. Bibliometric indicators are commonly known to have unconditional distributions that display overdispersion very often. However, IRT count data models deal with conditional distributions and the findings for Mutz and Daniel's dataset convincingly demonstrate that underdispersion at the item level can be present in the data and needs to be taken appropriately into account. Otherwise, technical problems and inaccurate estimation of the reliability of capacity estimates are expected. It is therefore highly recommended that a careful comparison of various competing IRT count data models precedes any examination of reliability.

Beyond the reliability concept as examined in this work (i.e., person separation of capacity as a latent variable) one might be interested in the reliability of a single observed indicator (see Allison, 1978). Dispersion parameters in the used IRT models in this work were found to be indicator-specific across both studied datasets, and they were estimated jointly with all other model parameters. However, when only a subset of indicators is available with the others missing at random, the derived model parameters can be used to calculate ability based on the subset and also the conditional variance and reliability. This can also be done in the extreme case of a single indicator (e.g., total number of published articles). This is conceptually similar, but mathematically distinct from Allison's (1978) approach, which is based on a conditional Poisson assumption for each indicator, and does not require parameter estimates from a whole set of indicators. These reliability considerations are beyond the scope of the current work and we strongly recommend future research that closely examines varying conceptions of measurement precision.

This study is clearly limited to less complex models as used by Mutz and Daniel (2018b). They used variants of the ICCPCM (Doebler et al. 2014) including a negative binomial extension of the model within a Bayesian estimation framework. We decided to use more parsimonious models in this work for several reasons. First, the focus was on the interplay between reliability of researcher capacity estimates and the used distributional models which is already quite complex. Hence, we chose simpler models for a better focus on our research question. Second, a ICCPCM extension to the CMPCM (i.e., an ICCMPCM) seems possible and straightforward in terms of theory, but estimation routines that allow application of such an extension are to the best of our knowledge currently not available. For example, Bayesian estimation of the CMPCM – as the basis for an

ICCCMPCM – is currently not available in ready-to-use software packages for GLMMs. Hence, given this might become an available alternative in the future, potential extensions of the CMPCM such as an ICCCMPCM will deserve a close examination.

Moreover, there are calls in the literature to use a variety of indicators (Moed and Halevi 2015) and others consider mere productivity as the most important indicator for funding and promotion (Yair and Goldstein 2020). In this work, productivity across careers of inventors and multiple bibliometric indices for a sample of social science researchers were studied, but the intention of this work was not to take any position of which approach to measurement (i.e., mere productivity vs. multi-faceted indicators) might be better. In fact, any evaluation of researchers' performance should be best guided by the concrete goals and consequences related to the respective selection decision (or other reason for such an evaluation).

Finally, it should be noted that recent research has extended the CMP distribution, especially with the goal to make the NB distribution one of the limit cases (Chakraborty and Ong 2016; Chakraborty and Imoto 2016; Imoto 2014). This is a potential avenue for a unification of the presented models, albeit regression modelling software including one of the extensions and random effects is currently not available. We also caution that none of the mentioned distributions lends itself naturally to a mean parametrization in the sense of Huang (2017), complicating model interpretation.

Conclusion

The current work has added important points to consider when such an evaluation is based on IRT-based estimates of researcher capacity. A sufficiently large set of alternative models has been examined with lower requirements with respect to sample size as compared to previous studies (Mutz and Daniel 2018b). Importantly, the CMPCM was introduced in this work that balances needed flexibility and sample size requirements. Most importantly, this study clearly demonstrated that model choice must precede any analysis of the reliability of researcher capacity estimates. Similarly, when researcher capacity standard errors are to be used, say to construct confidence intervals, model choice is equally important. Finally, count data models in general were shown to be well-suited for contexts in which decisions about the best researchers are to be made because all count data models displayed the highest level of conditional reliability at the top-ranges of researcher capacity.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *2nd International Symposium on Information Theory* 267–281. Akadémiai Kiadó
- Allison, P. D. (1978). The reliability of variables measured as the number of events in an interval of time. *Sociological Methodology*, 9, 238. <https://doi.org/10.2307/270811>.
- Alvarez, P., & Pulgarín, A. (1996). Application of the Rasch model to measuring the impact of scientific journals. *Publishing Research Quarterly*, 12(4), 57–64. <https://doi.org/10.1007/BF02680575>.
- Baghaei, P., Ravand, H., & Nadri, M. (2019). Is the d2 test of attention rasch scalable? analysis with the rasch poisson counts model. *Perceptual and Motor Skills*, 126(1), 70–86. <https://doi.org/10.1177/0031512518812183>.
- Barton, K. (2019). *MuMIn: Multi-Model Inference*. R package version 1.43.15. <https://CRAN.R-project.org/package=MuMIn>
- Bonat, W. H., Jørgensen, B., Kokonendji, C. C., Hinde, J., & Demétrio, C. G. B. (2018). Extended poisson–tweedie: Properties and regression models for count data. *Statistical Modelling: An International Journal*, 18(1), 24–49. <https://doi.org/10.1177/1471082X17715718>.
- Brooks, M, E, Kristensen, K, Benthem, K, J, van, Magnusson A, Berg CW, Nielsen A, Skaug, HJ, Mächler M, & Bolker, BM (2017) Glmm Balances Speed and Flexibility Among Packages for Zero inflated Generalized Linear Mixed Modeling. *The R Journal* 9(2), 378–390 <https://doi.org/10.32614/RJ-2017-066>
- Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Multivariate applications series. Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 307–333). Routledge/Taylor & Francis Group.
- Brown, Anna. (2018). Item Response Theory Approaches to Test Scoring and Evaluating the Score Accuracy. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing* (pp. 607–638). John Wiley & Sons, Ltd. <https://doi.org/https://doi.org/10.1002/9781118489772.ch20>
- Burrell, Q. L. (2007). Hirsch's h-index: A stochastic model. *Journal of Informetrics*, 1(1), 16–25. <https://doi.org/10.1016/j.joi.2006.07.001>.
- Chakraborty, S., & Ong, S. H. (2016). A COM-Poisson-type generalization of the negative binomial distribution. *Communications in Statistics - Theory and Methods*, 45(14), 4117–4135. <https://doi.org/10.1080/03610926.2014.917184>.
- Chakraborty, S., & Imoto, T. (2016). Extended Conway-Maxwell-Poisson distribution and its properties and applications. *Journal of Statistical Distributions and Applications*, 3(1), 5. <https://doi.org/10.1186/s40488-016-0044-1>.
- Cohen, J. E. (1981). Publication rate as a function of laboratory size in three biomedical research institutions. *Scientometrics*, 3(6), 467–487. <https://doi.org/10.1007/BF02017438>.
- Consul, P. C., & Famoye, F. (1992). Generalized poisson regression model. *Communications in Statistics - Theory and Methods*, 21(1), 89–109. <https://doi.org/10.1080/03610929208830766>.
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme package in R. *Journal of Statistical Software*. <https://doi.org/10.18637/jss.v039.i12>.
- Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics*, 7(4), 861–873. <https://doi.org/10.1016/j.joi.2013.08.006>.
- Doebler, A., Doebler, P., & Holling, H. (2014). A latent ability model for count data and application to processing speed. *Applied Psychological Measurement*, 38(8), 587–598. <https://doi.org/10.1177/0146621614543513>.
- Doebler, A., & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson Counts model. *Learning and Individual Differences*, 52, 121–128. <https://doi.org/10.1016/j.lindif.2015.01.013>.
- Faddy, M. J., & Bosch, R. J. (2001). Likelihood-based modeling and analysis of data underdispersed relative to the poisson distribution. *Biometrics*, 57(2), 620–624. <https://doi.org/10.1111/j.0006-341X.2001.00620.x>.
- Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, 11(1), 93–103. <https://doi.org/10.1177/014662168701100107>.
- Fischer, G. H. (1987). Applying the principles of specific objectivity and of generalizability to the measurement of change. *Psychometrika*, 52(4), 565–587. <https://doi.org/10.1007/BF02294820>.

- Forthmann, B., Çelik, P., Holling, H., Storme, M., & Lubart, T. (2018). Item response modeling of divergent-thinking tasks: A comparison of Rasch's Poisson model with a two-dimensional model extension. *International Journal of Creativity and Problem Solving*, 28(2), 83–95.
- Forthmann, B., Gerwig, A., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2016). The be-creative effect in divergent thinking: The interplay of instruction and object frequency. *Intelligence*, 57, 25–32. <https://doi.org/10.1016/j.intell.2016.03.005>.
- Forthmann, B., Günhe, D., & Doebl, P. (2019). Revisiting dispersion in count data item response theory models: The Conway–Maxwell–Poisson counts model. *British Journal of Mathematical and Statistical Psychology*. <https://doi.org/10.1111/bmsp.12184>.
- Forthmann, B., Szardenings, C., & Dumas, D. (2019). Testing equal odds in creativity research. *Psychology of Aesthetics, Creativity, and the Arts*. <https://doi.org/10.1037/aca0000294>.
- Glänzel, W., & Moed, H. F. (2013). Opinion paper: Thoughts and facts on bibliometric indicators. *Scientometrics*, 96(1), 381–394. <https://doi.org/10.1007/s11192-012-0898-z>.
- Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21(4), 347–360. <https://doi.org/10.1111/j.1745-3984.1984.tb01039.x>.
- Guikema, S. D., & Goffelt, J. P. (2008). A flexible count data regression model for risk analysis. *Risk Analysis*, 28(1), 213–223. <https://doi.org/10.1111/j.1539-6924.2008.01014.x>.
- Gulliksen, H. (1950). *Theory of Mental Tests*. John Wiley & Sons.
- Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools*. National Bureau of Economic Research. <https://www.nber.org/papers/w8498.pdf>
- Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed). Cambridge University Press.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572. <https://doi.org/10.1073/pnas.0507655102>.
- Holling, H., Böhning, W., & Böhning, D. (2015). The covariate-adjusted frequency plot for the Rasch Poisson counts model. *Thailand Statistician*, 13, 67–78.
- Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling: An International Journal*, 17(6), 359–380. <https://doi.org/10.1177/1471082X17697749>.
- Huber, J. C. (2000). A statistical analysis of special cases of creativity. *The Journal of Creative Behavior*, 34(3), 203–225. <https://doi.org/10.1002/j.2162-6057.2000.tb01212.x>.
- Huber, J. C., & Wagner-Döbler, R. (2001). Scientific production: A statistical analysis of authors in mathematical logic. *Scientometrics*, 50(2), 323–337. <https://doi.org/10.1023/A:1010581925357>.
- Huber, J. C., & Wagner-Döbler, R. (2001). Scientific production: A statistical analysis of authors in physics, 1800–1900. *Scientometrics*, 50(3), 437–453. <https://doi.org/10.1023/A:1010558714879>.
- Hung, L.-F. (2012). A negative binomial regression model for accuracy tests. *Applied Psychological Measurement*, 36(2), 88–103. <https://doi.org/10.1177/0146621611429548>.
- Imoto, T. (2014). A generalized Conway–Maxwell–Poisson distribution which includes the negative binomial distribution. *Applied Mathematics and Computation*, 247, 824–834. <https://doi.org/10.1016/j.amc.2014.09.052>.
- Jansen, M. G. H. (1995). The Rasch Poisson counts model for incomplete data: An application of the EM algorithm. *Applied Psychological Measurement*, 19(3), 291–302. <https://doi.org/10.1177/014662169501900307>.
- Jansen, M. G. H., & van Duijn, M. A. J. (1992). Extensions of Rasch's multiplicative poisson model. *Psychometrika*, 57(3), 405–414. <https://doi.org/10.1007/BF02295428>.
- Ketzler, R., & Zimmermann, K. F. (2013). A citation-analysis of economic research institutes. *Scientometrics*, 95(3), 1095–1112. <https://doi.org/10.1007/s11192-012-0850-2>.
- Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. L., et al. (2014). Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). *Research Policy*, 43(6), 941–955. <https://doi.org/10.1016/j.respol.2014.01.012>.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. L. Erlbaum Associates.
- Moed, H. F., & Halevi, G. (2015). Multidimensional assessment of scholarly research impact: The multidimensional assessment of scholarly research impact. *Journal of the Association for Information Science and Technology*, 66(10), 1988–2002. <https://doi.org/10.1002/asi.23314>.
- Mutz, R., & Daniel, H.-D. (2018a). The Bayesian Poisson Rasch model. *Data [Text/csv,0.007 MB] Research Collection*. <https://doi.org/10.3929/ETHZ-B-000271425>.

- Mutz, R., & Daniel, H.-D. (2018b). The bibliometric quotient (BQ), or how to measure a researcher's performance capacity: A Bayesian Poisson Rasch model. *Journal of Informetrics*, 12(4), 1282–1295. <https://doi.org/10.1016/j.joi.2018.10.006>.
- Ogasawara, H. (1996). Rasch's multiplicative poisson model with covariates. *Psychometrika*, 61(1), 73–92. <https://doi.org/10.1007/BF02296959>.
- R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rasch, G. (1960). *Studies in mathematical psychology: I*. Probabilistic models for some intelligence and attainment tests: Nielsen & Lydiche.
- Sahel, J.-A. (2011). Quality versus quantity: Assessing individual research performance. *Science Translational Medicine*, 3(84), 8413. <https://doi.org/10.1126/scitranslmed.3002249>.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>.
- Sellers, K. F., & Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics*, 4(2), 943–961. <https://doi.org/10.1214/09-AOAS306>.
- Sun, Y., & Xia, B. S. (2016). The scholarly communication of economic knowledge: A citation analysis of google scholar. *Scientometrics*, 109(3), 1965–1978. <https://doi.org/10.1007/s11192-016-2140-x>.
- Wang, W.-C. (1999). Direct estimation of correlations among latent traits within IRT framework. *Methods of Psychological Research Online*, 4(2), 47–68.
- Wright BD, & Masters, GN (1982) *Rating scale analysis*. Mesa Press.
- Yair, G., & Goldstein, K. (2020). The Annus Mirabilis paper: Years of peak productivity in scientific careers. *Scientometrics*, 124(2), 887–902. <https://doi.org/10.1007/s11192-020-03544-z>.
- Zhang, F., Bai, X., & Lee, I. (2019). Author impact: Evaluations, predictions, and challenges. *IEEE Access*, 7, 38657–38669. <https://doi.org/10.1109/ACCESS.2019.2905955>.
- Zhu, W., & Safrit, M. J. (1993). The calibration of a sit-ups task using the rasch poisson counts model. *Canadian Journal of Applied Physiology*, 18(2), 207–219. <https://doi.org/10.1139/h93-017>.