Scientometrics (2023) 128:4753–4764
https://doi.org/10.1007/s11192-023-04756-9

Researcher capacity estimation based on the *Q* model: a generalized linear mixed model perspective

Boris Forthmann[1]

Received: 9 January 2023 / Accepted: 24 May 2023 / Published online: 5 June 2023

Abstract

Chance models of scientific creative productivity allow estimation of researcher capacity. One prominent such model is the *Q* model in which the impact of a scholarly work is modeled as a multiplicative function of researcher capacity and a potential impact (i.e., luck) parameter. Previous work estimated researcher capacity based on an approximation of the *Q* parameter. In this work, however, I outline how the *Q* model can be estimated within the framework of generalized linear mixed models. This way estimates of researcher capacity (and all other parameters of the *Q* model) are readily available and obtained by standard statistical software packages. Usage of such software further allows comparing different distributional assumptions and calculation of reliability of the *Q* parameter (i.e., researcher capacity). This is illustrated for a large dataset of multidisciplinary scientists ($N = 20{,}296$). The Poisson *Q* model was found to have negligibly better predictive accuracy than the original normal *Q* model. Reliability estimates revealed excellent reliability of *Q* estimates with conditional reliability being mostly in acceptable ranges. Reliability of *Q* parameter estimates further depended heavily on the number of publications of a scientist with reliability increasing with the number of papers. The future and limitations of the *Q* model in the context of researcher capacity estimation are thoroughly discussed.

Keywords *Q* model · Researcher capacity · Generalized linear mixed model

Mathematics Subject Classification 62P25

JEL Classification C18

Introduction

Scientists must produce something to be recognized by their peers. Typically, scientists produce scholarly articles and then counting how often such articles are cited by researchers is used as a proxy for impact (i.e., more citations imply more impact in a field; Hartley, 2017; Pan & Fortunato, 2014). But how are scientific productivity and the impact of the

✉ Boris Forthmann
boris.forthmann@wwu.de

[1] Institute of Psychology, University of Münster, Münster, Germany

produced works related to each other? Which role play luck and a researcher's capability in a scientific career? Theoretical models that provide answers to such questions take into account randomness (i.e., luck) as well as individual differences (Simonton, 2004, 2010; Sinatra et al., 2016). For example, the Q model has been proposed to decompose variation in impact of scholarly publications into researcher capacity and luck (Sinatra et al., 2016). The Q model was theoretically developed and empirically tested to increase the understanding of the interplay between productivity (i.e., number of publications) and impact in scientific careers with the more distal goal that such an understanding is needed for more practical questions of research evaluation (Sinatra et al., 2016). In this work, I pick up this aim and investigate the Q model's capability as a vehicle for research evaluation at the individual level within a well-established statistical framework.

The Q model is a multiplicative model in which impact of a paper is predicted by the product of the researcher capacity parameter Q and the potential impact parameter p_α. Both parameters are further examined in relation to researcher productivity. Statistical inference and estimation of related model parameters is based on a trivariate log-normal model. However, given that impact or other bibliometric indicators for the measurement of researcher capacity are count data, the current work embeds the Q model into the comprehensive statistical framework of Generalized Linear Mixed Models (GLMMs; e.g., Stroup, 2013). GLMMs allow choices between several count data distributions with a mean parameterization that preserves the architecture of the Q model, but are expected to provide a much better fit to the data. Thus, alternative distributional assumptions for the Q model can be made and empirically tested. Importantly, conceptualizing the Q model within a GLMM framework comes along with several more advantages: (a) wide accessibility of the model because it can be estimated by means of well-known GLMM statistical packages, (b) direct estimates of the Q parameters (and other model parameters) can be obtained (i.e., no approximate formula is required), and c) overall and specific reliability estimates can be quantified.

The *Q* model

The impact $S_{i\alpha}$ of paper α written by scientist i is modeled as a multiplicative function of researcher capacity parameter Q_i and the potential impact p_α. A paper's impact is commonly approximated by citation counts with $S_{i\alpha} \in \mathbb{N}_0$. However, citation counts might be replaced by rating counts of movies or books and play counts of songs when the Q model is estimated in other domains such as the movie, book, or music industries (Janosov et al., 2020), yet still count data are used. In the context of the Q model, researcher capacity refers to the capability to utilize the existing knowledge base in a way that enhances the impact of a paper (Sinatra et al., 2016). Notably, other researchers have used a similar conceptualization of individual differences in research performance (Mutz & Daniel, 2018, 2019). Later Janosov et al. (2020) tested the Q model across multiple domains (e.g., movies, music, and book writing) and referred more broadly to the ability to consistently produce high-impact work. However, for simplicity I will use the term researcher capacity throughout this work. Like the Q_i parameter, the potential impact parameter p_α is unobserved an represents the impact a paper might potentially have in the future. Hence, according to Sinatra et al. (2016), a high-impact scholarly paper will most likely result when a researcher with high Q_i works (by chance) on a paper project with high potential impact p_α. Specifically, the Q model is then defined by the following equation (Janosov et al., 2020; Sinatra et al., 2016)

$$S_{i\alpha} = Q_i p_\alpha. \tag{1}$$

The Q model implies further a log-linear model with additive components $\log(Q_i) = \widehat{Q}_i$ and $\log(p_\alpha) = \widehat{p}_\alpha$. The additive nature of the Q model at the log-level facilitates to see that the luck component is actually a residual term that captures everything in a paper's impact that cannot be explained by a person main effect. Thus, technically the luck component comprises of paper main effects, researcher-paper interaction effects, random noise, as well as sampling error. However, given that scholarly papers are uniquely nested within researchers, these different effects on impact cannot be separated in the context of the Q model.

Within the Q model framework, a trivariate normal distribution for $\widehat{Q}_i$, $\widehat{p}_\alpha$, and $\log(N_i) = \widehat{N}_i$ (i.e., the log of the number of publications published by scientist i) is assumed with zero correlation between $\widehat{p}_\alpha$ and both other parameters. In addition, also the correlation between $\widehat{Q}_i$ and $\widehat{N}_i$ is proposed to be zero (Janosov et al., 2020; Sinatra et al., 2016). Based on these model properties the Q model can be understood as implying a "random impact rule", i.e. the probability for the highest impact paper within a scientist's career is uniform. The Q model was found to empirically fit data of scientists (Janosov et al., 2020; Sinatra et al., 2016), as well as data of movie directors, pop musicians, and book authors (Janosov et al., 2020).

Estimation of $\widehat{Q}_i$ parameters were obtained by means of a two-step approach in Sinatra et al. (2016). First, the means and the covariance matrix of their trivariate normal distribution formulation of the Q model were estimated by maximizing the corresponding log-likelihood function. Specifically, they used the fmincon function from Matlab's Optimization package. The optimization function was run 100 times (ten times for each of the ten different starting conditions; cf. the supplemental material of Sinatra et al., 2016). The final estimates were the averages across all 100 runs of the fmincon function. Then, a maximum of the log-likelihood for exact calculation of the $\widehat{Q}_i$ estimates is obtained. The formula for approximate $\widehat{Q}_i$ as the average of a researchers log-impact across all papers was then derived based on this exact formula. In Janosov et al. (2020) a different algorithm (i.e., a covariance matrix adaptation evolution strategy algorithm) was used to find the parameters of the trivariate normal distribution. For calculation of individual $\widehat{Q}_i$ estimates, however, the approximation was used. Especially in comparison to the approach used in the first paper on the Q model by Sinatra et al. (2016) which involved running an optimization function for hundred times, it is expected that relying on proven estimation algorithms such as those implemented in widely used GLMM software will decrease the amount of computation time. I argue that this is one aspect of GLMM-based Q model estimation that has the potential to make the model more widely accessible for researchers.

The *Q* model as a GLMM

For the context of this work, the following formulation of the Q model will be used

$$\log\left(S_{i\alpha}\right) = \eta_{i\alpha} = \mu + \hat{Q}_i + \hat{p}_\alpha, \tag{2}$$

with linear predictor $\eta_{i\alpha}$ and an overall log-level intercept μ. It is further assumed that $\widehat{Q}_i \sim N(0, \sigma^2_{\widehat{Q}})$ and $\widehat{p}_\alpha \sim N(0, \sigma^2_{\widehat{p}})$. That is, both parameters are considered to be latent variables and explicitly treated as being uncorrelated, i.e. they are modeled as uncorrelated random effects. In addition, two different distributional assumptions are made in this work and empirically compared: (a) $S_{i\alpha}|\hat{Q}_i, \hat{p}_\alpha \sim N\left(\exp\left(\eta_{i\alpha}\right), \sigma_{\in}^2\right)$, and (b) $S_{i\alpha}|\hat{Q}_i, \hat{p}_\alpha \sim Poi(\exp(\eta_{i\alpha}))$. The normal Q model is well in accordance with previous work by Sinatra and colleagues (Janosov et al., 2020; Sinatra et al., 2016), whereas the Poisson Q model has not been implemented before. Indeed, using a genuine count distribution instead of approximating count data by a continuous distribution seems rather reasonable and has been done previously in related scientometric investigations (e.g., Mutz & Daniel, 2019).

First, the residual variance σ_ϵ^2 is fixed to zero to identify the model based on the normal distribution (i.e., residual variance is forced into the random luck parameter), i.e., all variance is modeled at the latent level and observations of $S_{i\alpha}$ are perfectly explained by $\hat{Q}_i$ and $\hat{p}_\alpha$. Furthermore, latent means for both latent variables would not be uniquely identified and hence only one overall log-level intercept μ is part of the Q model as a GLMM. This is different as compared to other estimation routines of the Q model (Janosov et al., 2020; Sinatra et al., 2016) in which latent means for both parameters are estimated. However, these two estimates are expected to add up to the overall μ parameter estimated in this work. For example, Sinatra et al. (2016) found latent means of 0.92 and 0.93 for potential impact and researcher capacity, respectively. Thus, the Q model formulation in this work would be expected to result in $\mu = 1.85$ for their data. Critically, estimation of researcher capacity or the potential impact parameter will not be influenced by this difference.

Quantifying measurement precision of $\hat{Q}_i$ estimates

Obtaining $\hat{Q}_i$ estimates can be useful for research and practical assessment contexts (e.g., research evaluation in the context of personnel selection). Yet, when using them researchers and stakeholders should be informed about the quality of these estimates. One aspect of the quality of $\hat{Q}_i$ estimates is measurement precision. Given that $\hat{Q}_i$ estimates can be obtained in GLMMs as maximum a posteriori estimates along with standard errors, it is possible to adopt the idea of empirical reliability from the psychometrics literature (Brown & Croudace, 2015; Green et al., 1984). Empirical (marginal) reliability is widely used in educational assessment (Forthmann et al., 2022; McNeish & Dumas, 2018) and has also been used in scientometric research (Forthmann & Doebler, 2021). Conceptually, empirical reliability estimates the squared correlation of the $\hat{Q}_i$ estimates with their true values. A value of one implies perfect reliability (i.e., maximum measurement precision) which means unity correlation between $\hat{Q}_i$ estimates and their true values, whereas a reliability of zero implies no shared variance between $\hat{Q}_i$ estimates and their true values (i.e., complete lack of measurement precision). Empirical reliability can be estimated by the following formula

$$\mathrm{Rel}\left(\hat{Q}\right) = 1 - \frac{\overline{SE}_{\hat{Q}}^2}{s_{\hat{Q}}^2}, \tag{3}$$

with $s_{\hat{Q}}^2$ being the estimated variance of the researcher capacity distribution and $\overline{SE}_{\hat{Q}}^2$ being the average squared standard error for the $\hat{Q}_i$ estimates (Brown & Croudace, 2015). Importantly, in practical assessment contexts one is not only interested in a general estimate of reliability. Instead, reliability of individual $\hat{Q}_i$ estimates would be of utmost importance. This is especially the case when the amount of available information—i.e., the number of papers for a researcher—is sparse. This is indeed not unlikely in personnel selection in academia when the applicant pool includes early career researchers, for example. Quantifying uncertainty of $\widehat{Q}_i$ estimates to better guide decision making is possible by means of conditional reliability estimates (i.e., reliability for a certain level of researcher capacity)

$$\mathrm{Rel}\left(\hat{Q}_i\right) = 1 - \frac{SE^2_{\hat{Q}_i}}{s^2_{\hat{Q}}}, \tag{4}$$

with $SE^2_{\hat{Q}_i}$ being the squared standard error of the researcher capacity estimate of the ith researcher. Finally, it should be noted that reliability of approximate $\widehat{Q}_i$ can be obtained by means of multilevel reliability (Snijders & Bosker, 2011) which is also often used in educational assessment contexts (Forthmann et al., 2022; Schatschneider et al., 2008).

Aim of the current study

The goal of this work is to fit Sinatra et al.'s (2016) Q model by means of standard GLMM software. This way the original normal Q model was evaluated against an alternative Poisson Q model. In addition, multiple questions related to researcher capacity as reflected by the $\widehat{Q}_i$ estimates were thoroughly investigated. First, the correlation between the $\widehat{Q}_i$ estimates and the log of the number of publications was examined. Second, empirical reliability estimates (Brown & Croudace, 2015) for the $\widehat{Q}_i$ and $\widehat{p}_\alpha$ estimates were obtained. Third, conditional reliability at all estimated levels of researcher capacity $\widehat{Q}_i$ and its relationship with the number of publications was assessed. Fourth, the correlations between $\widehat{Q}_i$ estimates based on different distributional assumptions were examined (the same was checked for $\widehat{p}_\alpha$ estimates too). Fifth, I correlated the original Q parameter approximation (Janosov et al., 2020; Sinatra et al., 2016) on the log-level with the estimates obtained in this study (to apply the approximation I omitted the mean of the luck distribution which does not affect correlations here). Finally, I obtained Janosov et al.'s R index which is the percentage of variance attributable to the $\widehat{p}_\alpha$ parameter (Janosov et al., 2020).

Method

Dataset

In this work I reanalyzed a dataset comprised of multi-disciplinary scientists (Liu et al., 2018). These data were made openly available by Liu et al. (2018) in a github repository: https://lu-liu.github.io/hotstreaks/. A total of $N = 20{,}296$ scientists was reanalyzed, yet as compared to the original paper I did not restrict the data to the subset with at least fifteen publications and a career length of minimum 20 years. This way it was possible to investigate reliability of researcher capacity estimates with rather few papers, for example. The dataset includes for each scientist the respective papers and how often they were cited within a 10-year time window. For more details on how the dataset was constructed the original paper (Liu et al., 2018) can be consulted. For the purpose of the current work, citation counts 10 years after publication were modeled as the dependent variable in the Q model. In addition, the (log-)number of each scientist's publications was analyzed.

Data analysis

The models were estimated by means of the package glmmTMB (Brooks et al., 2017) which allows flexible generalized linear modeling within the statistical software R (R Core Team, 2022). To prevent technical problems with the log-link combined with the normal distribution in case of zero citations, a value of one was added to the citation counts of each paper.[1] Maximum a posteriori estimates and their standard errors of the $\widehat{Q}_i$ and $\widehat{p}_\alpha$ parameters were obtained and used for further examination (e.g., calculation of empirical and conditional reliability).

Given that the normal Q model is based on a continuous distribution and the Poisson Q model on a discrete distribution, it is not possible to use likelihood-based information criteria such as the Akaike information criterion (Akaike, 1998) cannot be used. This is because the densities of both distributions are defined with respect to different measures (i.e., a Lebesgue measure for the continuous distribution and a counting measure for the discrete distribution; Commenges et al., 2015; Proust-Lima et al., 2012). Hence, I compared both models based on the following cross-validation approach: the available data of each researcher were sampled into two equal halves (a training dataset and a test dataset), both Q model variants were estimated for the training dataset, finally the μ and $\widehat{Q}_i$ estimates were used to predict the $\log\left(S_{i\alpha}\right)$ values in the test dataset. The predictive accuracy of both Q models was evaluated by means of root mean square error (RMSE), mean absolute error (MAE), and the Pearson correlation coefficient. Three researchers in the dataset had only one paper published and were excluded from the cross-validation. It should further be noted that any types of transformations imposed on the training data should also be part of the cross-validation. Hence, I subtracted a value of one from the predictions made based on the estimates obtained from the training dataset. The following R packages were used for the cross-validation analysis: tidyverse (Wickham et al., 2019) and caret (Kuhn, 2021).

Results

The parameter estimates of both models are depicted in Table 1. The estimated standard deviation of $\widehat{Q}_i$ was highly similar across both models, whereas the standard deviation for $\widehat{p}_\alpha$ was clearly higher for the normal Q model (even the confidence intervals for the same estimated parameter obtained from both models did not overlap). Also, the general intercept parameter differed between both models with a higher estimate resulting for the Poisson Q model (again confidence intervals did not overlap; see Table 1). In addition, the correlation between the $\widehat{Q}_i$ estimates and the log-transformed number of publications was found to be small for both models. The correlations were significantly larger than zero which is strictly speaking a violation of the Q model. Yet, this slight deviation might be negligible (see Sinatra et al., 2016, for a similar argument).

Next, reliability of the latent variables was evaluated. Reliability of the $\widehat{Q}_i$ estimates was excellent regardless of the model (see Table 1). Thus, the $\widehat{Q}_i$ estimates obtained from

[1] The Poisson Q model did not require this transformation, yet for better comparability it was employed for both investigated Q model variants. Naturally, the Poisson distribution can handle zeros and even excessive occurrences of zeros can be modeled by zero-inflated Poisson models. This idea was brought up by one anonymous reviewer and complementary results exploring zero-inflation can be found in an online supplementary file: https://osf.io/ga8wt/.

Table 1 Model estimates for both variants of the Q model

	Normal model		Poisson model	
	Estimate	95%-CI	Estimate	95%-CI
$s_{\widehat{Q}}$	0.68	[0.67, 0.68]	0.68	[0.67, 0.69]
$s_{\widehat{p}}$	1.34	[1.33, 1.34]	1.24	[1.24, 1.24]
μ	2.26	[2.25, 2.27]	2.32	[2.31, 2.33]
$Cor(\widehat{Q}_i, \widehat{N}_i)$	0.13	[0.12, 0.15]	0.15	[0.14, 0.16]
$Rel(\widehat{Q}_i)$	0.90		0.91	
$Rel(\widehat{p}_\alpha)$	0.98		0.89	
Janosov et al.'s R index	0.80		0.77	
Cross-validation				
RMSE	96.09		95.67	
MAE	22.99		22.88	
Pearson r	0.22	[0.22, 0.22]	0.23	[0.22, 0.23]

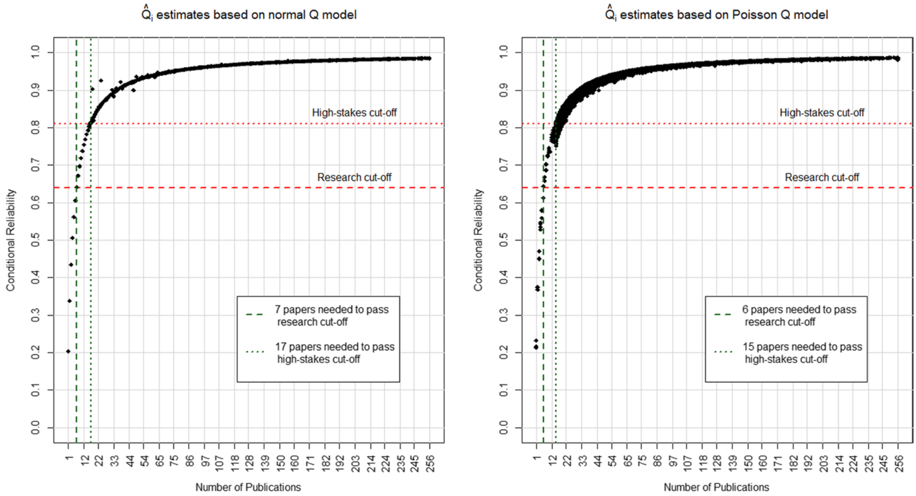

Fig. 1 Relationship between a researcher's number of publications (x axis) and reliability of the $\widehat{Q}_i$ estimate. Left: $\widehat{Q}_i$ estimates are based on the normal Q model. Right: $\widehat{Q}_i$ estimates are based on the Poisson Q model. Horizontal red lines refer to common reliability requirements (Ferrando & Lorenzo-Seva, 2018). Vertical dark green lines refer to the minimum number of papers needed to pass one of the two reliability cut-offs

both Q model variants were on average clearly reliable enough for high-stakes contexts of research assessment (Ferrando & Lorenzo-Seva, 2018), whereas multilevel reliability of approximate $\widehat{Q}_i$ which was estimated to be 0.88 did not pass this cut-off. In addition, the random luck component was more reliably estimated with the normal Q model as compared to the Poisson Q model. Conditional reliability (i.e., reliability of single $\widehat{Q}_i$ estimates) was examined as a function of the number of papers for $\widehat{Q}_i$ estimates based on the normal Q model (see left in Fig. 1), and $\widehat{Q}_i$ estimates based on the Poisson Q model (see right in Fig. 1). Clearly, reliability of $\widehat{Q}_i$ estimates increases as a nonlinear function of the number of papers. Figure 1 further illustrates that reliability of $\widehat{Q}_i$ estimates was sufficient for research purposes (i.e., reliability > 0.64) when a researcher had at least seven (normal Q model) or six (Poisson Q model) papers published, whereas seventeen (normal Q model) or fifteen (Poisson Q model) published papers were needed for reaching reliability required for high-stakes research assessment contexts (i.e., reliability > 0.81).

In addition, a very strong correlation between $\widehat{Q}_i$ estimates obtained from both model variants was close to unity with $r = 0.997$, 95% CI [0.996, 0.997]. A comparably high correlation was found for the $\widehat{p}_\alpha$ estimates obtained from both models with $r = 0.993$, 95% CI [0.993, 0.993]. The correlation between $\widehat{Q}_i$ estimates as obtained by the normal Q model and approximate $\widehat{Q}_i$ was also close to unity with $r = 0.998$, 95% CI [0.998, 0.998] (same for the Poisson Q model: $r = 0.995$, 95% CI [0.994, 0.995]). This latter finding emphasizes the validity of Q model estimation within the GLMM framework. Finally, Janosov et al.'s R index (Janosov et al., 2020) was evaluated. The values obtained for both models were again highly comparable and they suggest that around 80% of the variation in log-transformed units are attributable to luck rather than researcher capacity.

Finally, the cross-validation revealed that the predictive accuracy of the Poisson Q model was negligibly better for RMSE, MAE, and Pearson r as compared to the normal Q model (see Table 1). The $\widehat{Q}_i$ estimates obtained from the training dataset correlated also close to unity with the approximate $\widehat{Q}_i$ from the test dataset (normal Q model: $r = 0.96$, 95% CI [0.96, 0.96]; Poisson Q model: $r = 0.96$, 95% CI [0.96, 0.96]).

Discussion

Sinatra et al.'s (2016) Q model is an influential theoretical model of scientific productivity that implies, for example, a random impact rule (cf. Simonton, 2010). The Q model conceptualizes the impact of a scholarly paper as a multiplicative function of a researcher capacity parameter and a paper's inherent potential (including luck). This suggests that the Q model might have merit for research assessment contexts, yet up to now no widely accessible tools for estimating the parameters (in particular the $\widehat{Q}_i$ estimates) were known. Looking at the model from the perspective of GLMMs (Stroup, 2013) as done in this work was aimed at filling this gap. As a GLMM the Q model parameters can be estimated with any statistical software that includes GLMM estimation functions. In this work, I relied on the R package glmmTMB (Brooks et al., 2017) which has a formula argument for the dispersion model. By means of this functionality it was possible to estimate the normal Q model with a residual variance of zero. This might hint interested researchers at functionalities needed, when one wishes to switch to a different software package. To further facilitate such endeavors, all code to reproduce the reported findings (https://osf.io/ga8wt/) and the dataset used (https://lu-liu.github.io/hotstreaks/) are openly available.

In addition, the current work extends previous findings by Sinatra et al. (2016) and Janosov et al. (2020) in important ways: a) the Q model was evaluated and tested for a different dataset, b) the Q model was empirically tested against a highly competitive alternative distributional assumption, c) reliability as a critical assessment property was comprehensively evaluated, and d) evidence of validity for Q model estimation within a GLMM framework was provided. First, the Q model proposes that the $\widehat{Q}_i$ estimates and the log-transformed number of papers $\widehat{N}_i$ are uncorrelated. Sinatra et al. (2016) reported covariance estimates that imply a moderate correlation of $r = 0.34$ between $\widehat{Q}_i$ estimates and $\log(N_i)$, yet they interpreted this level of covariation as a "slight association" (Sinatra et al., 2016, p. 599).

In this work, however, the correlation was found to be much smaller and, hence, evidence for Q model fit to the data in this work was stronger as compared to previous work. Second, given that the raw citation counts in the dataset were count data, it was expected that a genuine count data distribution such as the Poisson should be competitive for the originally formulated Q model based on the normal distribution. Indeed, the Poisson Q model had better predictive accuracy and displayed better reliability of the $\widehat{Q}_i$ estimates. Many findings, however, were highly comparable across both distributional assumptions and $\widehat{Q}_i$ estimates taken from both Q model variants correlated close to unity.

The fact that $\widehat{Q}_i$ estimates were found to be highly reliable in this work makes the Q model highly attractive for a variety of research assessment contexts. Regardless of the underlying distributional assumption used, reliability estimates clearly surpassed the required level for high-stakes assessment contexts. In this vein, the current work contributes to current discussions and evaluated approaches to researcher capacity estimation by means of item response theory (Alvarez & Pulgarín, 1996; Forthmann & Doebler, 2021; Mutz & Daniel, 2018), for example. Furthermore, it was clearly demonstrated that reliability of individual researcher capacity estimates depends heavily and non-linearly on the number of published papers. Indeed, this relationship between published papers and measurement precision is known for quite some time (Cronbach, 1941; Dennis, 1958), yet it has not yet been discussed in relation to the Q model. Only researchers who published already a certain number of papers may be accurately evaluated by means of the Q model. In the current work, six or seven papers would be needed for a reliability level required for research purposes while fifteen or seventeen papers would be needed for high-stakes assessment purposes. It is easy to imagine promising candidates for an academic position who have not yet published at least fifteen papers and evaluating their researcher capacity will be associated with greater uncertainty as compared to more productive candidates. Finally, it should be noted that reliability of approximate $\widehat{Q}_i$ values did not pass the typical cut-off for high-stakes evaluation, which emphasizes a somewhat better measurement quality of the maximum a posteriori estimates obtained from the GLMMs for practical purposes. Admittedly, such cut-offs are always somewhat arbitrary and local committees may decide if the approximate values or estimates obtained from a fitted GLMM Q model are sufficiently reliable for a given assessment context. Still, such cut-offs have heuristic value and are helpful to communicate decisions and their associated uncertainty to stakeholders. This is very important information if one wishes to transfer the Q model to practical assessment contexts.

Furthermore, it is notable that the estimates of researcher capacity based on the GLMM version of the Q model and the estimates obtained by Sinatra et al.'s (2016) approximation formula correlate close to unity. This emphasizes the validity of the estimation approach proposed and evaluated in the current work. To further validate the approach, one could think of simulation studies as a nice future endeavor. Such a simulation could, for example, focus on much smaller sample sizes as studied here and in prior other work on the Q model (Janosov et al., 2020; Sinatra et al., 2016). This would greatly help to see how well the Q model might work in more practical assessment settings in which far less data points (e.g., the pool of applicants for an academic position) as compared to the current work are available.

I further tried to fit the Poisson Q model with additional zero-inflation parameter (as suggested by one anonymous reviewer) which did not converge for technical reasons. Hence, I re-estimated the Poisson model with only the researcher capacity parameter (i.e., the luck parameter was omitted; cf. Mutz & Daniel, 2019) and compared this with a model that additionally incorporated the probability for an excess zero. The latter model displayed better fit and the probability for an excess zero was estimated as 16%. In addition, the correlation of the $\widehat{Q}_i$ estimates from the zero-inflated Poisson model and the $\widehat{Q}_i$ approximation was $r=0.81$. However, partially this comparably small correlation was also a result of leaving out the luck parameter. That is, Poisson models including only researcher capacity displayed a similar (yet somewhat stronger) correlation with the approximate $\widehat{Q}_i$ ($r=0.90$). Thus, overall, it seems that omitting the luck parameter played a more important role in reducing the correlation (please note that the correlations between $\widehat{Q}_i$ estimates from the complete models and the approximate $\widehat{Q}_i$ were close to unity) and modeling of zero-inflation seemed to have only a small effect here which decreased this correlation only slightly more. Clearly, more research is needed to reveal the potential of explicitly modeling zero-inflation in the context of the Q model.

Finally, there are notable similarities between the Poisson Q model and Mutz and Daniel's multi-membership Poisson model (Mutz & Daniel, 2019). In fact, Mutz and Daniel develop their model based on a much simpler Poisson model that includes an intercept parameter and a researcher capacity random effect. Hence, their initial model is identical to the Poisson Q model in which the random luck parameter is omitted. They further extend their model by a reference value for field normalization and a weighting of the person random effect. The weighting is chosen according to a specific scheme for fractional counting (e.g., all co-authors receive equal weight or the first author receives the highest weight for contributing to a scholarly paper). Based on this model individual differences in researcher capacity estimates are controlled for the expected citations in a field (i.e., comparisons of researchers across fields becomes possible in a fair manner). In addition, it is taken into account that scholarly papers often have more than one author with potentially unequal contributions by each author. It seems straightforward to simply add a luck parameter to their model for a field normalized multi-membership Poisson Q model (or a log-normal variant analogous to the original Q model formulation). Importantly, fitting this model would not be possible for the dataset used in this work as the publications of the authors were not coded by a unique identifier. Hence, a careful empirical investigation of such an extended Q model is a promising avenue for future research.

Overall, the reported findings look promising for the Q model to become a flexible tool for research and practice. While the current work has direct implications for practical research assessment (e.g., the Q parameter can be highly reliably estimated), there are more far reaching consequences for the Q model as well. For example, the GLMM perspective would easily allow to further integrate covariates on either the researcher or publication level to explain impact of scholarly work. In addition, one might think of a multidimensional extension of the Q model that allows modeling of several quality dimensions at the same time. Future work is needed to explore these options for the Q model, but the path is well laid out from here onwards.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

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

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., & Bolker, B. M. (2017). GlmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal, 9*(2), 378. 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.

Commenges, D., Proust-Lima, C., Samieri, C., & Liquet, B. (2015). A universal approximate cross-validation criterion for regular risk functions. *The International Journal of Biostatistics*. https://doi.org/10.1515/ijb-2015-0004

Cronbach, L. J. (1941). The reliability of ratio scores. *Educational and Psychological Measurement, 1*(1), 269–277. https://doi.org/10.1177/001316444100100121

Dennis, W. (1958). The age decrement in outstanding scientific contributions: Fact or artifact? *American Psychologist, 13*(8), 457–460. https://doi.org/10.1037/h0048673

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78*(5), 762–780. https://doi.org/10.1177/0013164417719308

Forthmann, B., & Doebler, P. (2021). Reliability of researcher capacity estimates and count data dispersion: A comparison of Poisson, negative binomial, and Conway-Maxwell-Poisson models. *Scientometrics, 126*(4), 3337–3354. https://doi.org/10.1007/s11192-021-03864-8

Forthmann, B., Förster, N., & Souvignier, E. (2022). Multilevel and empirical reliability estimates of learning growth: A simulation study and empirical illustration. *Frontiers in Education, 7*, 920704. https://doi.org/10.3389/feduc.2022.920704

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

Hartley, J. (2017). Authors and their citations: A point of view. *Scientometrics, 110*(2), 1081–1084. https://doi.org/10.1007/s11192-016-2211-z

Janosov, M., Battiston, F., & Sinatra, R. (2020). Success and luck in creative careers. *EPJ Data Science, 9*(1), 9. https://doi.org/10.1140/epjds/s13688-020-00227-w

Kuhn, M. (2021). *caret: Classification and regression training*. R package version 6.0–90. https://CRAN.R-project.org/package=caret

Liu, L., Wang, Y., Sinatra, R., Giles, C. L., Song, C., & Wang, D. (2018). Hot streaks in artistic, cultural, and scientific careers. *Nature, 559*(7714), 396–399. https://doi.org/10.1038/s41586-018-0315-8

McNeish, D., & Dumas, D. (2018). Calculating conditional reliability for dynamic measurement model capacity estimates: DMM reliability. *Journal of Educational Measurement, 55*(4), 614–634. https://doi.org/10.1111/jedm.12195

Mutz, R., & Daniel, H.-D. (2018). 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

Mutz, R., & Daniel, H.-D. (2019). How to consider fractional counting and field normalization in the statistical modeling of bibliometric data: A multilevel Poisson regression approach. *Journal of Informetrics, 13*(2), 643–657. https://doi.org/10.1016/j.joi.2019.03.007

Pan, R. K., & Fortunato, S. (2014). Author impact factor: Tracking the dynamics of individual scientific impact. *Scientific Reports, 4*(1), 4880. https://doi.org/10.1038/srep04880

Proust-Lima, C., Amieva, H., & Jacqmin-Gadda, H. (2012). Analysis of multivariate mixed longitudinal data: A flexible latent process approach. *British Journal of Mathematical and Statistical Psychology, n/a-n/a.* https://doi.org/10.1111/bmsp.12000

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Schatschneider, C., Wagner, R. K., & Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences, 18*(3), 308–315. https://doi.org/10.1016/j.lindif.2008.04.005

Simonton, D. K. (2004). *Creativity in science: Chance, logic, genius, and zeitgeist* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139165358

Simonton, D. K. (2010). Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity. *Physics of Life Reviews, 7*(2), 156–179. https://doi.org/10.1016/j.plrev.2010.02.002

Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science, 354*(6312), aaf5239. https://doi.org/10.1126/science.aaf5239

Snijders, T. A. B., & Bosker, R. (Eds.). (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Stroup, W. W. (2013). *Generalized linear mixed models: Modern concepts, methods and applications*. CRC Press, Taylor & Francis Group.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686