


Applying Negative Binomial Distribution in Diagnostic Classification Models for Analyzing Count Data

Applied Psychological Measurement
2023, Vol. 47(1) 64–75
© The Author(s) 2022



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01466216221124604
journals.sagepub.com/home/apm


Ren Liu¹ , Ihnwhi Heo¹ , Haiyan Liu¹, Dexin Shi², and Zhehan Jiang³

Abstract

Diagnostic classification models (DCMs) have been used to classify examinees into groups based on their possession status of a set of latent traits. In addition to traditional item-based scoring approaches, examinees may be scored based on their completion of a series of small and similar tasks. Those scores are usually considered as count variables. To model count scores, this study proposes a new class of DCMs that uses the negative binomial distribution at its core. We explained the proposed model framework and demonstrated its use through an operational example. Simulation studies were conducted to evaluate the performance of the proposed model and compare it with the Poisson-based DCM.

Keywords

diagnostic classification model, negative binomial, Poisson, count data

In operational tests, examinees may be scored based on their completion of a series of small and similar tasks. For example, students are asked to read out loud 50 words; children are asked to memorize the sequence of 10 animals; patients are asked to select from a list of 20 symptoms. If the purpose of those tests were to classify examinees as master/non-masters of a group of abilities or bearer/non-bearer of a group of disorders, diagnostic classification models (DCMs) could be candidate scoring models. When responses on those types of tasks are scored, they are usually treated as count variables. For the 50 words that students read out loud, we typically do not treat

¹University of California, Merced, CA, USA

²University of South Carolina, Columbia, SC, USA

³Peking University, Beijing, China

Corresponding Authors:

Ren Liu, Quantitative Methods, Measurement, and Statistics (QMMS), University of California, Merced, Merced, CA 95343, USA.

Email: rliu45@ucmerced.edu

Zhehan Jiang, Institute of Medical Education & National Center for Health Professions Education Development, Peking University, Beijing 100191, China.

Email: jiangzhehan@gmail.com

them as 50 items because they are quite similar. They were typically given to students, for example, in five blocks of 10 words each. When students' responses are scored, we count how many words in each block they answer correctly. For this example, each student would get five counts, one for each block. Those five counts are used to estimate their latent trait characteristics. If DCMs were applied to score count data, one could use the Poisson-based DCM (PDCM; Liu et al., 2021) that has recently been proposed. A feature of the PDCM, or any statistical model that uses a Poisson distribution, is that the mean and variance have been fixed to be equal. This constraint may be unnecessary and sometimes unrealistic given that the variance is often greater than the mean. To relax this constraint, this study proposes a more flexible DCM framework for scoring count data that comes from a series of small, and sometimes repetitive tasks. The proposed framework uses the negative binomial distribution at its core, which allows the mean and variance of the count variable to be separately estimated. In the next section, we first introduce the necessary theoretical foundations before presenting our proposed modeling framework.

Theoretical Framework

Diagnostic Classification Models

Diagnostic classification models are multidimensional confirmatory latent class models. They could be appropriate scoring models when a researcher's primary goal is to classify examinees into pre-defined groups (aka latent classes). To form those groups, one needs to first hypothesize at least two latent traits (commonly known as attributes), and treat them as categorical variables, which can be either binary (0 or 1, representing non-possession and possession of attributes) or polytomous (e.g., 0, 1, 2, representing non-possession, partial possession, and full possession of attributes). For K attributes, the combination of attribute possession groups forms 2^K possible latent classes. Example applications of DCMs include obtaining student reading skill profiles (e.g., George & Robitzsch, 2021; Jang et al., 2013), obtaining student math and science skill profiles (e.g., Kabiri et al., 2017; Lee et al., 2011), classifying people with personality types (e.g., Liu & Shi, 2020; Xi et al., 2020), and diagnosing mental disorders for patients (e.g., de la Torre et al., 2018; Templin & Henson, 2006).

Although many DCMs have been proposed with different parameterizations of the measurement component and/or the structural component, we can use a general form of DCM as an example to introduce the specification of a DCM. When item responses are scored in a binary fashion, the general form of DCMs is the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009). Let $\alpha_c = \{\alpha_1, \dots, \alpha_K\}$ index 2^K latent classes that contain different combinations of the K attributes. The LCDM defines the probability of examinees with latent class c scoring a "1" on item i as

$$P(X_i = 1 | \alpha_c) = \frac{\exp[\omega_{0,i} + \omega_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i)]}{1 + \exp[\omega_{0,i} + \omega_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i)]} \quad (1)$$

where $\omega_{0,i}$ is the intercept, $\omega_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i) = \sum_{k=1}^K \omega_{1,i,k}(\alpha_{c,k} q_{i,k}) + \sum_{k=1}^{K-1} \sum_{k'=k+1}^K \omega_{2,i,k,k'}(\alpha_{c,k} \alpha_{c,k'} q_{i,k} q_{i,k'}) + \dots + \omega_{K,i,1,\dots,K} \prod_{k=1}^K (\alpha_{c,k} q_{i,k})$ including all main effects and interaction effects, and \mathbf{q}_i is a vector of 1s and 0s indicating whether item i measures each attribute. The \mathbf{q}_i is usually shown in an $I \times K$ matrix commonly known as a Q -matrix. The core component of the LCDM is the $\omega_{0,i} + \omega_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i)$, on which constraints can be added to arrive at most of the earlier DCMs such as the deterministic inputs, noisy, and gate" (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001), and the linear logistic model (LLM; Maris, 1999). The DINA model can be seen as the LCDM with only intercepts and the highest-order interaction terms, and the LLM can be seen as

the LCDM with only intercepts and main effects. By integrating the Poisson distribution into the core component of the LCDM, Liu et al. (2021) developed the PDCM.

Poisson Distribution and the PDCM Framework

The Poisson distribution is commonly used to analyze count variables. It expresses the probability that the number of events (s) in a defined time interval as

$$P(X = s) = \frac{\lambda^s}{s! \exp^{\lambda}}, \lambda > 0 \quad (2)$$

where λ is the rate parameter, representing both the mean and variance of X .

The framework of PDCMs uses the Poisson distribution at their core, which expresses the probability of examinees with a given latent class obtaining a count score of s on item i as

$$P(X_i = s | \mathbf{a}_c) = \frac{\lambda_{ci}^s}{s! \exp^{\lambda_{ci}}}, \quad (3)$$

where $\lambda_{ci} = \omega_{0,i} \times \omega_i^T \mathbf{h}(\mathbf{a}_c, \mathbf{q}_i)$. Comparing Equations 1–3, we can see that the framework of PDCMs is structured similarly to the probability mass function of the Poisson distribution while using the core component of the LCDM. The count score s refers to the number of correct or incorrect responses in one item block. For example, if an examinee gets 9 letters correct in a block of 10 letters on a letter recognition exam, their $s = 9$ for that item block if one chooses to model correct counts, or $s = 1$ for that item block if one chooses to model incorrect counts.

Negative Binomial Distribution and Its Use in Item Response Theory Models

When we use Poisson-based models to analyze count variables, the data must conform to equidispersion where the mean and variance are equal because there is only one parameter λ . However, variance is often greater than the mean in an operational dataset. For example, the average amount of incorrectness in 10 letter recognition tasks may be one, but the variance could be greater than one. As an alternative to the Poisson distribution, the negative binomial distribution is more flexible by allowing the mean and variance to be different. The probability mass function of the negative binomial distribution can be expressed as

$$P(X = s) = \binom{s + \phi - 1}{s} \left(\frac{\mu}{\mu + \phi} \right)^s \left(\frac{\phi}{\mu + \phi} \right)^{\phi} \quad (4)$$

where μ is the mean parameter, and ϕ is the parameter that controls dispersion so that the variance can be written as $\mu + \frac{\mu^2}{\phi}$. The negative binomial distribution has shown its merit in dealing with overdispersion in the fields of psychology (e.g., Gardner et al., 1995), sociology (e.g., Land et al., 1996), ecology (e.g., Ver Hoef & Boveng, 2007), and medicine (e.g., Miaou, 1994).

In the item response theory (IRT) framework, the negative binomial distribution has also been applied in studies such as Hung (2012), and Magnus and Thissen (2017). Following the parameterization used in Equation (3), the one-parameter form of a negative binomial IRT model can be written as

$$P(X_i = s | \theta_e) = \left(\frac{\Gamma(s + r_i^{-1})}{\Gamma(s + 1)\Gamma(r_i^{-1})} \right) \left(\frac{\exp(\theta_e - b_i)}{r_i^{-1} + \exp(\theta_e - b_i)} \right)^{s_i} \left(\frac{r_i^{-1}}{r_i^{-1} + \exp(\theta_e - b_i)} \right)^{r_i^{-1}} \quad (5)$$

where θ_e represents examinee e 's latent trait, b_i represents the difficulty for item i , and r_i^{-1} is the dispersion parameter equivalent to the ϕ in equation (4). We could see that equation (5). is a direct extension from equation (4). where the μ is specified as $\exp(\theta_e - b_i)$.

A Framework of Count Diagnostic Classification Models

The framework of Count Diagnostic Classification Models (CDCMs) is developed based on 1) applying the negative binomial distribution to the general form of DCMs in a similar fashion to how it is applied to IRT models; and 2) the relationship between the general form of DCMs and its special cases.

Like the PDCM framework, we extract the core of the LCDM: $\omega_{0,i} + \omega_i^T \mathbf{h}(\mathbf{a}_c, \mathbf{q}_i)$ to replace the μ in equation (4). as the mean. As a result, the general form of the CDCMs can be written as

$$P(X_i=s|\mathbf{a}_c) = \left(\frac{\Gamma(s+r_i^{-1})}{\Gamma(s+1)\Gamma(r_i^{-1})} \right) \left(\frac{\exp[\omega_{0,i} + \omega_i^T \mathbf{h}(\mathbf{a}_c, \mathbf{q}_i)]}{r_i^{-1} + \exp[\omega_{0,i} + \omega_i^T \mathbf{h}(\mathbf{a}_c, \mathbf{q}_i)]} \right)^{s_i} \left(\frac{r_i^{-1}}{r_i^{-1} + \exp[\omega_{0,i} + \omega_i^T \mathbf{h}(\mathbf{a}_c, \mathbf{q}_i)]} \right)^{r_i^{-1}} \quad (6)$$

Comparing equation (6) to equation (5), we can see that the negative binomial distribution is applied to the DCMs in a similar fashion to that in IRT models where we replaced $\exp(\theta_e - b_i)$ with $\exp[\omega_{0,i} + \omega_i^T \mathbf{h}(\mathbf{a}_c, \mathbf{q}_i)]$. To ensure that possessing more attributes does not decrease the probability of having a higher score, the main effect and interaction effect parameters are all constrained to be non-negative. Similar to equation (5), r_i^{-1} controls for the variance where a positive value signifies that the variance is greater than the mean. The identification of the proposed CDCM requires considerations in combining the identification requirements of a general DCM and the special “item block” property of the CDCM. Gu and Xu (2019) and Xu and Zhang (2016) have demonstrated that the identifiability of a DCM only depends on the Q -matrix structure. Following their framework, two necessary conditions are required to identify the proposed CDCM. First, the Q -matrix needs to be “complete” (Chiu et al., 2009), meaning that it can differentiate all attribute profiles. Second, each attribute needs to be associated with at least three item blocks. The number of tasks within each item block is not much of a concern. Note that the condition of three item blocks is a necessary but not sufficient condition for identification because one needs to consider the degrees of freedom based on the different number of attribute profiles and loadings in the Q -matrix.

After deriving the general form of the CDCM framework from the LCDM, we can re-parameterize the $\omega_{0,i} + \omega_i^T \mathbf{h}(\mathbf{a}_c, \mathbf{q}_i)$ component to obtain other subsumed models for count data. For example, to form a DINA-type model, we replace the $\omega_{0,i} + \omega_i^T \mathbf{h}(\mathbf{a}_c, \mathbf{q}_i)$ in the LCDM with an intercept $\omega_{0,i}$ and an effect $\omega_{1,i} \prod_{k=1}^K \alpha_{c,k}^{q_{i,k}}$ that signifies all the effects of all related attributes. The resulting CDCM-DINA can be written as

$$P(X_i=s|\mathbf{a}_c) = \left(\frac{\Gamma(s+r_i^{-1})}{\Gamma(s+1)\Gamma(r_i^{-1})} \right) \left(\frac{\exp[\omega_{0,i} + \omega_{1,i} \prod_{k=1}^K \alpha_{c,k}^{q_{i,k}}]}{r_i^{-1} + \exp[\omega_{0,i} + \omega_{1,i} \prod_{k=1}^K \alpha_{c,k}^{q_{i,k}}]} \right)^{s_i} \left(\frac{r_i^{-1}}{r_i^{-1} + \exp[\omega_{0,i} + \omega_{1,i} \prod_{k=1}^K \alpha_{c,k}^{q_{i,k}}]} \right)^{r_i^{-1}} \quad (7)$$

Similarly, if we replace the $\omega_{0,i} + \omega_{1,i} \prod_{k=1}^K \alpha_{c,k}^{q_{i,k}}$ from the CDCM-DINA with an intercept $\omega_{0,i}$ and the main effects of all related attributes $\sum_{k=1}^K \omega_{1,i,k}(\alpha_{c,k} q_{i,k})$, we can express the CDCM-LLM as

$$P(X_i = s | \alpha_c) = \left(\frac{\Gamma(s + r_i^{-1})}{\Gamma(s + 1) \Gamma(r_i^{-1})} \right) \left(\frac{\exp[\omega_{0,i} + \sum_{k=1}^K \omega_{1,i,k}(\alpha_{c,k} q_{i,k})]}{r_i^{-1} + \exp[\omega_{0,i} + \sum_{k=1}^K \omega_{1,i,k}(\alpha_{c,k} q_{i,k})]} \right)^{s_i} \left(\frac{r_i^{-1}}{r_i^{-1} + \exp[\omega_{0,i} + \sum_{k=1}^K \omega_{1,i,k}(\alpha_{c,k} q_{i,k})]} \right)^{r_i^{-1}} \quad (8)$$

Through these two examples, we hope readers can see how other DCMs may also be formulated within the CDCM framework to analyze count data.

Operational Study

We aim to achieve two purposes through conducting this operational study. The first is to demonstrate the use of the CDCM, and the second is to compare its performance with the PDCM. The dataset was used in [Liu et al. \(2021\)](#) which contains 808 examinees' responses to 24 item blocks that measure three attributes. Item blocks 1–8 measure number recognition, item blocks 9–16 measure color recognition, and item blocks 17–24 measure object recognition. As a result, a simple-structure Q -matrix was developed based on the content specification. In the Q -matrix, each item measures one attribute. In each item block, there are 10 numbers, colors, or objects, meaning that the maximum count score in each block (aka item) is 10. A preliminary review of the dataset shows that most examinees got most of the items correct. Therefore, we chose to model the incorrect count in each item. If other researchers choose to do this in their studies, they want to make sure that the interpretation of the attribute possession status (α_k) aligns with the direction of the number of counts (whether it is a count of incorrect answers or correct answers).

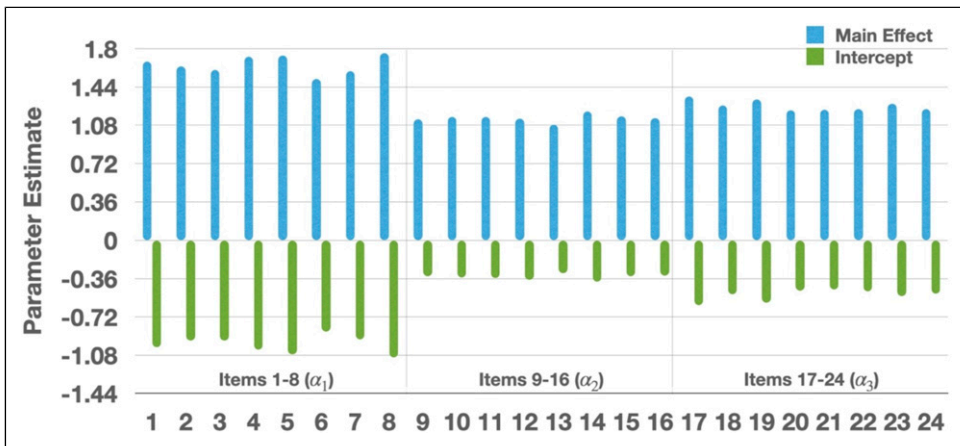


Figure 1. Mean of the posterior distribution for the intercept ($\omega_{0,i}$) and main effect ($\omega_{1,i}$) parameters.

Fitting the CDCM

We used the Stan (Carpenter et al., 2017) program for parameter estimation. The Stan code used to estimate the CDCM is shared in the [Supplemental Appendix](#). In the measurement model, we estimated 24 intercepts ($\omega_{0,i}$), 24 main effects ($\omega_{1,i}$), and 24 dispersion parameters (r_i^{-1}). Similar to Ghosh et al. (2018) and Liu et al. (2021), we implemented the following priors for the above-mentioned parameters: $\omega_{0,i} \sim \text{Normal}(0, 2)$, $\tilde{\omega}_{1,i} \sim \text{Normal}(0, 2)$, $r_i \sim \text{Cauchy}(0, 5)$. We ran four chains with a length of 20,000, where parameters were sampled from the last 10,000 draws. The chains converged with Gelman–Rubin’s \hat{R} values (Gelman & Rubin, 1992) all close to 1.00. The general CDCM fit the data well with posterior predictive p -values (Gelman et al., 2013) at 0.54.

Mean of the posterior distribution for each parameter was used as a point estimate and displayed in Figure 1. All the intercepts were below 0 and all the main effects were (constrained to be) above 0. Typically, we say that an item is good when it has high discrimination where small intercepts and large main effects help differentiate between those who possess and not possess an attribute. The parameter estimates show that items in this dataset are of good quality. Items 1–8 (measuring α_1) had larger main effects and smaller intercepts than items 17–24 (measuring α_3), which also had larger main effects and smaller intercepts than items 9–16 (measuring α_2). Figure 2 shows the standard deviation of the posterior distribution of the parameters. Generally, parameters with larger absolute mean values had larger standard deviations in their distributions, but all of them were below 0.12. Results for the r_i parameters are shown in Figure 3. We can see all the r_i values were positive, meaning that the variance of examinees’ responses on each item is greater than the mean. Item 4 had the largest r value: 0.012, which is still very small. Small r_i values (i.e., large r_i^{-1} values) lead to a large denominator of the equation for variance $\left(\mu + \frac{\mu^2}{r_i^{-1}}\right)$, which means that the difference between the variance and mean was small for each item.

Comparing the Results Between the CDCM and the PDCM

If the difference between the variance and the mean was small enough under the CDCM, we could also fit the PDCM to the dataset and examine the differences. Therefore, we also fit the PDCM to the dataset, using the same Stan specifications. Although parameter estimates are not directly comparable, we compared model fit and classification agreement between the CDCM and the PDCM. Regarding relative model fit, we computed the leave-one-out cross-validation information criterion (LOOIC; Vehtari et al., 2017), where smaller values indicate better fit. The LOOIC values for the CDCM and the PDCM were 72.3 and 72.5, respectively. Using the standard errors of the difference in their expected predictive accuracy estimates, we found that the two models did not fit

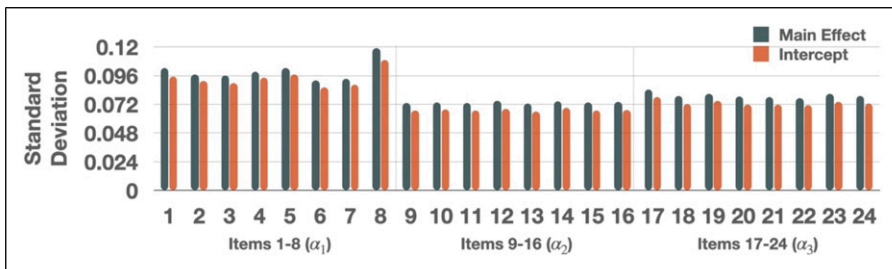


Figure 2. Standard Deviation of the posterior distribution for the intercept ($\omega_{0,i}$) and main effect ($\omega_{1,i}$) parameters.

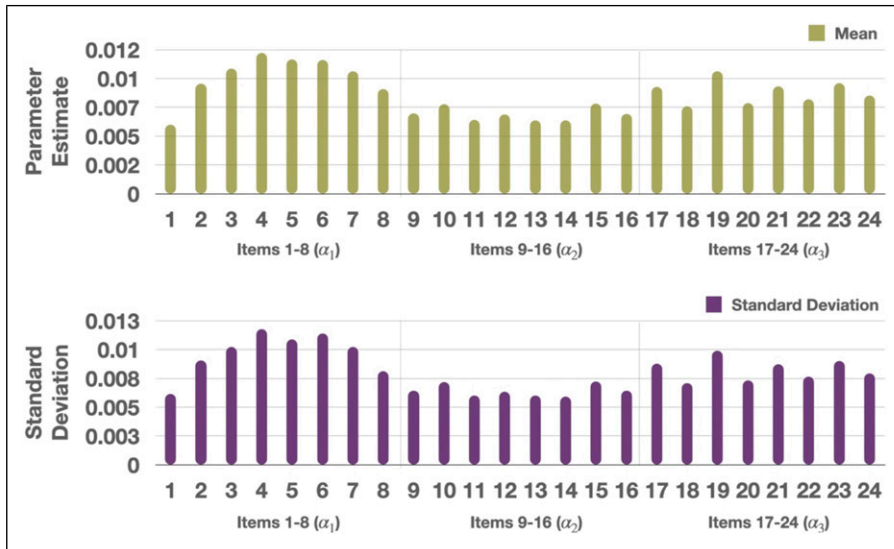


Figure 3. Mean and standard deviation of the posterior distribution for the inverse of the dispersion parameter (r_i).

significantly differently. Consider that the PDCM is more restrictive than the CDCM, the results suggest that the item-level variance of examinees' responses was not much different from the mean. This lack-of-difference is supported by the small r_i values that we mentioned in the previous section. Regarding classification agreement, results show that every single examinee (i.e., 100%) was classified with the same attribute profile. In other words, if a researcher fit both the CDCM and the PDCM to this dataset, they would get identical results on examinee scores.

But what if there were relatively larger differences between the mean and the variance (i.e., larger r_i values)? How would that affect the results of fitting the PDCM and the CDCM? We aim to explore that through a simulation study in the next section.

Simulation Study

The above operational study demonstrated similar performance of the PDCM and the CDCM when the data dispersion is small with r_i values close to 0. The purpose of the current simulation study is two-fold: 1) to examine whether the CDCM could produce unbiased parameter estimates; and 2) to investigate whether the CDCM performs better than the PDCM under different degrees of data dispersion. The simulation study is couched in the operational study setting where we used the parameters obtained from the operational study to represent real data situations.

Data Generation

To simulate 808 examinees' responses to 24 item blocks, we used the mean of each parameter's posterior distribution that we obtained through the CDCM in the simulation study as the true parameter values. We varied the dispersion parameter at two levels: small and large. For small dispersion conditions, we sampled r^{-1} from *Uniform* (0.8, 1.0). We chose this range to mimic the operational dataset used in Hung (2012). Since the mean of the distribution for each parameter was close to one in our dataset, using $r^{-1} = 1$, the standard deviation of responses in each item block

Table 1. Mean Bias for the Count Diagnostic Classification Model Parameters in the Simulation Study.

Item	Intercept		Main Effect	
	Small r	Large r	Small r	Large r
1	0.092	0.101	0.003	−0.003
2	0.119	0.126	−0.048	−0.051
3	0.124	0.136	−0.059	−0.069
4	0.104	0.116	−0.044	−0.053
5	0.136	0.150	−0.047	−0.059
6	0.102	0.110	−0.059	−0.064
7	0.076	0.085	0.021	0.015
8	0.142	0.148	−0.023	−0.025
9	0.076	0.099	0.045	0.033
10	0.074	0.099	0.058	0.043
11	0.044	0.055	0.106	0.110
12	0.036	0.054	0.073	0.067
13	0.056	0.086	0.020	−0.005
14	0.039	0.057	0.064	0.059
15	0.044	0.060	0.088	0.086
16	0.028	0.036	0.097	0.107
17	0.105	0.120	0.009	0.002
18	0.051	0.062	0.061	0.060
19	0.067	0.084	0.039	0.029
20	0.056	0.068	0.065	0.061
21	0.085	0.098	0.032	0.026
22	0.072	0.078	0.037	0.041
23	0.101	0.117	0.018	0.009
24	0.090	0.109	0.063	0.050

would be around 1.41. This represents conditions with small dispersion. For large dispersion conditions, we sampled r^{-1} from *Uniform* (0.2, 0.4). We chose this range to mimic the operational dataset used in Magnus and Thissen (2017). Using a mean of one and $r^{-1} = 0.2$, the standard deviation of responses in each item block would be around 2.45. This represents conditions with a large dispersion. Using the above specifications, we generated 40 datasets using R (R Core Team, 2019), where 20 used small dispersion parameters and 20 used large dispersion parameters. Both the CDCM and the PDCM were fitted to each dataset.

Results 1: Parameter Recovery of the CDCM

To examine parameter recovery for the CDCM, we computed the mean bias and RMSE for each parameter across iterations and listed them in Tables 1 and 2. Overall, the average bias for the intercept parameters were 0.080 and 0.094 under the small and large dispersion conditions, respectively. The average bias for the main effect parameters were 0.026 and 0.020 under the two conditions, respectively. We also examined the accuracy of examinees’ attribute classifications. The average attribute-wise classification accuracy was 0.984 and 0.981, for the small and large dispersion conditions, respectively. The profile-wise classification accuracy was 0.848 and 0.835, for the two conditions, respectively.

Table 2. Mean RMSE for the Count Diagnostic Classification Model Parameters in the Simulation Study.

Item	Intercept		Main Effect	
	Small <i>r</i>	Large <i>r</i>	Small <i>r</i>	Large <i>r</i>
1	0.146	0.151	0.134	0.136
2	0.162	0.166	0.144	0.146
3	0.155	0.164	0.147	0.151
4	0.159	0.166	0.125	0.130
5	0.177	0.188	0.150	0.155
6	0.163	0.166	0.149	0.151
7	0.133	0.136	0.119	0.117
8	0.172	0.177	0.105	0.106
9	0.111	0.126	0.108	0.111
10	0.106	0.122	0.122	0.120
11	0.112	0.117	0.165	0.171
12	0.093	0.100	0.164	0.164
13	0.107	0.121	0.115	0.115
14	0.086	0.097	0.153	0.155
15	0.089	0.103	0.155	0.169
16	0.071	0.072	0.153	0.169
17	0.133	0.145	0.127	0.128
18	0.116	0.114	0.126	0.119
19	0.125	0.134	0.138	0.140
20	0.093	0.101	0.109	0.108
21	0.130	0.138	0.132	0.134
22	0.125	0.128	0.137	0.142
23	0.137	0.151	0.122	0.133
24	0.126	0.140	0.135	0.134

Table 3. Classification Accuracy of the Count Diagnostic Classification Model and the Poisson-Based Diagnostic Classification Model in the Simulation Study.

	Count diagnostic classification model		Poisson-based diagnostic classification model	
	Small <i>r</i>	Large <i>r</i>	Small <i>r</i>	Large <i>r</i>
α_1	0.985	0.984	0.822	0.819
α_2	0.993	0.991	0.835	0.823
α_3	0.974	0.969	0.779	0.766
Profile	0.848	0.835	0.600	0.587

Results 2: Comparing the CDCM to the PDCM

We computed the classification accuracy of the CDCM and the PDCM for each dataset and presented the results in Table 3. Overall, model selection has a larger impact than the change in the relative data dispersion. In the operational study, we observed identical classification results between the CDCM and the PDCM when the item-level *r* values were estimated to be very small. In the simulation study, *r* values were larger than those in the operational study, and the PDCM

produced much worse classification results compared to the CDCM. Between the two levels of r values, the classification results were very similar. Recall from our data generation, small and large r values correspond to a standard deviation of around 1.41 and 2.45, respectively. In other words, the PDCM was very sensitive to mean-variance mismatch, and it produced undesirable classification results even when the standard deviation was 0.4 above a mean of 1.

Discussion

Although measurement models with the negative binomial distribution involved are less commonly seen than those with the Poisson distribution, they are more flexible and may provide more accurate parameter estimates if mean and variance are not equal.

The proposed CDCM framework may be used when: 1) multiple attributes are being tested through a series of similar tasks, and 2) the purpose of the test is to classify examinees as mastery/non-mastery or possession/non-possession of those attributes.

In that situation, we first recommend computing the mean and variance of the responses on each item block. This could be an easy first step to get a sense of what we could expect for the mean-variance differences. Then we recommend fitting both the PDCM and the CDCM to the dataset and comparing model fit and classification agreement. If one obtains small r_i estimates like the operational study, they may also fit the PDCM (which forces the mean and variance to be equal) and see if the CDCM fits significantly better. If the CDCM does not fit significantly better, it should produce very similar, if not identical results compared to the PDCM. If the CDCM fit significantly better and the classification agreement between the two models is low, one may consider proceeding with then more flexible CDCM. When r_i was at least not too small, the simulation study demonstrated that the CDCM produced more accurate classification results than the PDCM.

Although the CDCM seems more complicated than the PDCM, the number of parameters being estimated is about the same between the two models. As demonstrated in the model formulation, the number of CDCM parameters equals the number of PDCM parameters plus the number of item blocks (one r for each item block). If there were 10 item blocks, the CDCM has 10 more parameters. In many situations, examinees' responses to each item block that measure the same attribute are very similar, we could also consider fixing the r_i values to be the same across all items (i.e., using r to replace r_i). In our operational example, the parameter estimates for the r_i values were very similar, and it would make sense to consider adding this constraint. Studies such as [Hung \(2012\)](#) had this similar constraint imposed in negative binomial IRT models. When the number of items is large, this constraint may help reduce the number of unnecessary parameters.

For future research, the following directions may be considered. First, one could investigate the effects of different prior distributions for r_i . The priors for r_i were sampled from a Cauchy distribution in our study. The traditional prior choice would be an inverse gamma prior. Between inverse gamma and Cauchy, we chose Cauchy in our example mainly because [Gelman \(2006\)](#) and [Polson and Scott \(2012\)](#) argued that Cauchy performed better than inverse gamma. Specifically, [Gelman \(2006\)](#) argued that if the variance estimates are very small (e.g., close to 0), inverse gamma may be sensitive to inference problems. Although Cauchy is thick-tailed similar to inverse gamma, Cauchy is even less informative compared to inverse gamma. Second, one could explore other models that could also deal with the mean-variance mismatch in real data. For example, the Conway–Maxwell–Poisson counts model ([Conway & Maxwell, 1962](#)) is a candidate model that could deal with both underdispersion and overdispersion. Third, response time could be considered and jointly modeled with the count data. With the advancement in new technology, tracking examinees' response process and gathering such data has been easier. Joint modeling could provide more accurate latent trait estimation as well as more informative feedback. We hope

this line of research on using DCMs to analyze item responses from a series of simple tasks could benefit operational testing practice.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Ren Liu  <https://orcid.org/0000-0002-6708-4996>

Ihnwhi Heo  <https://orcid.org/0000-0002-6123-3639>

Supplemental Material

Supplemental material for this article is available online.

References

- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Chiu, C. Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74(4), 633–665. <https://doi.org/10.1007/s11336-009-9125-0>
- Conway, R. W., & Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12(2), 132–136.
- de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, 51(4), 281–296. <https://doi.org/10.1080/07481756.2017.1327286>
- Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, 118(3), 392–404. <https://doi.org/10.1037/0033-2909.118.3.392>
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3), 515–534. <https://doi.org/10.1214/06-ba117a>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC Press. (Ch. 6).
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–511. <https://doi.org/10.1214/ss/1177011136>
- George, A. C., & Robitzsch, A. (2021). Validating theoretical assumptions about reading with cognitive diagnosis models. *International Journal of Testing*, 21(2), 105–129. <https://doi.org/10.1080/15305058.2021.1931238>
- Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, 13(2), 359–383. <https://doi.org/10.1214/17-ba1051>
- Gu, Y., & Xu, G. (2019). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika*, 84(2), 468–483. <https://doi.org/10.1007/s11336-018-9619-8>
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26(4), 301–321. <https://doi.org/10.1111/j.1745-3984.1989.tb00336.x>

- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2), 191–210. <https://doi.org/10.1007/s11336-008-9089-5>
- 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>
- Jang, E. E., Dunlop, M., Wagner, M., Kim, Y. H., & Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. *Language Learning*, 63(3), 400–436. <https://doi.org/10.1111/lang.12016>
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272. <https://doi.org/10.1177/01466210122032064>
- Kabiri, M., Ghazi-Tabatabaei, M., Bazargan, A., Shokoohi-Yekta, M., & Kharrazi, K. (2017). Diagnosing competency mastery in science: An application of GDM to TIMSS 2011 data. *Applied Measurement in Education*, 30(1), 27–38. <https://doi.org/10.1080/08957347.2016.1258407>
- Land, K. C., McCall, P. L., & Nagin, D. S. (1996). A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models: With empirical applications to criminal careers data. *Sociological Methods & Research*, 24(4), 387–442. <https://doi.org/10.1177/0049124196024004001>
- Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, 11(2), 144–177. <https://doi.org/10.1080/15305058.2010.534571>
- Liu, R., Liu, H., Shi, D., & Jiang, Z. (2021). Poisson diagnostic classification models: A framework and an exploratory example. *Educational and Psychological Measurement*, 82(3), 506–516. <https://doi.org/10.1177/00131644211017961>
- Liu, R., & Shi, D. (2020). Using diagnostic classification models in psychological rating scales. *The Quantitative Methods for Psychology*, 16(5), 442–456. <https://doi.org/10.20982/tqmp.16.5.p442>
- Magnus, B. E., & Thissen, D. (2017). Item response modeling of multivariate count data with zero inflation, maximum inflation, and heaping. *Journal of Educational and Behavioral Statistics*, 42(5), 531–558. <https://doi.org/10.3102/1076998617694878>
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64(2), 187–212. <https://doi.org/10.1007/bf02294535>
- Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, 26(4), 471–482. [https://doi.org/10.1016/0001-4575\(94\)90038-8](https://doi.org/10.1016/0001-4575(94)90038-8)
- Polson, N. G., & Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4), 887–902. <https://doi.org/10.1214/12-ba730>
- R Core Team (2019). R (version 3.6): R Foundation for Statistical Computing [Computer Software].
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11(3), 287–305. <https://doi.org/10.1037/1082-989X.11.3.287>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88(11), 2766–2772. <https://doi.org/10.1890/07-0043.1>
- Xi, C., Cai, Y., Peng, S., Lian, J., & Tu, D. (2020). A diagnostic classification version of Schizotypal Personality Questionnaire using diagnostic classification models. *International Journal of Methods in Psychiatric Research*, 29(1), Article e1807. <https://doi.org/10.1002/mpr.1807>
- Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, 81(3), 625–649. <https://doi.org/10.1007/s11336-015-9471-z>