

Psychometrically skewed distributions lead to pseudoclustering
when using mixture models:

A commentary on “Heterogeneity in children
at risk of math learning difficulties” (Munez et al., 2023)

Enrico Toffalini

Department of General Psychology, University of Padova, Italy

March 30, 2024

Abstract

Munez et al. (2023) challenge the dimensional framework on learning difficulties by suggesting that children with math difficulties might cluster into qualitatively different types, rather than representing a single homogeneous population. They appropriately employ mixture modeling as the analytical framework of choice. However, the use of non-normally distributed sum scores from mathematical tasks as indicators makes inference problematic. While non-normality is not uncommon in psychological data, it poses a violation of assumptions in the mixture models employed, potentially inflating the number of detected latent classes. Monte Carlo simulations demonstrate that Munez et al. (2023) may have had a heightened likelihood of detecting multiple clusters even in the absence of true latent classes. This case underscores a broader issue in clustering individuals based on psychological data, emphasizing the importance of conducting a priori sanity checks before employing clustering for inference.

Psychometrically skewed distributions lead to pseudoclustering when using mixture models: A commentary on “Heterogeneity in children at risk of math learning difficulties” (Munez et al., 2023)

In a recent study, Muñoz, Bull, Lee, and Ruiz (2023) challenge the dimensional framework on learning difficulties, suggesting that children with math learning difficulties (MLD) might cluster into two (or even three) qualitatively different types, rather than representing cases of a single homogeneous population. They fit a series of confirmatory factor analyses (CFA), latent profile analyses (LPA), and factor mixture models. The authors adequately identified mixture models as an ideal method for investigating the existence of subpopulations within a larger set of cases, thus unveiling hidden heterogeneity in what is otherwise traditionally considered an undifferentiated condition. In a similar fashion, in their recent comprehensive review on the “transdiagnostic revolution,” Astle, Holmes, Kievit, and Gathercole (2022) suggested using clustering methods as a primary means of discovering true underlying neurocognitive types beneath the surface of traditional diagnostic categories in neurodevelopmental disorders. While such a data-driven approach should not be used in isolation, Astle et al. (2022) present it as a significant complement to the toolbox of the researcher investigating neurodiversity, alongside dimensional methods.

”Clustering techniques, including those based on mixture models, rely on certain assumptions to ensure valid statistical inference. Like other statistical methods, when these assumptions are not met, results might still be valid, but their validity is not necessarily guaranteed. Many mixture models used in psychology assume multivariate normal distributions. Unfortunately, the normality assumption is highly unlikely to be met in our discipline (Micceri, 1989). Moreover, even when assumptions are fully met, results might be incorrect. For instance, a lack of statistical power may imply that some parameters are not adequately modeled. If the ground truth is unknown, however, such risks remain concealed. Luckily, data simulation allows for conducting a priori sanity checks to quantify these risks beforehand. In the case of clustering techniques, one could simulate data from a given multivariate distribution, with pre-specified features such as correlation, skewness, and kurtosis coefficients (based on a priori knowledge or on one’s own data). If data are drawn from a single population but the chosen clustering method predominantly favors solutions that involve multiple clusters, this is clearly a problem, suggesting that another method should be used.

An example of the aforementioned is provided by Toffalini, Girardi, Giofrè, and Altoè (2022). The authors demonstrated that Gaussian mixture models, under different specific conditions, tend to incorrectly detect multiple clusters despite data being drawn from a single population. Surpris-

ingly, this occurred even when the distributional assumption was met (i.e., data were generated from perfectly Gaussian multivariate distributions). Such an inflation of detected clusters arose when the sample size was insufficient for modeling the existing covariance across indicators within the detected clusters (e.g., the sample size was medium or even large, but the correlations across multiple indicators were weak). Remarkably, under specific combinations of sample size, correlation coefficients, and number of indicators, it was virtually guaranteed that Gaussian mixture models would detect multiple latent classes/clusters where none existed. Another challenge is that for true clusters to be correctly detected, their separation across clustering indicators must be sufficiently large (i.e., at least Cohen’s $d = 0.8$ on several independent dimensions; see Tein, Coxe, & Cham, 2013) to make it unlikely that they could have remained undetected until data-driven analysis was conducted. Such large effect sizes on many orthogonal variables may even seem implausible in psychological research itself (Toffalini et al., 2022).

In Muñoz et al. (2023), the separations between the two detected clusters are quite substantial: the intercepts differ by more than 1.0 SD in all three indicators of math ability, and several other domain-general and domain-specific cognitive variables exhibit between-cluster mean differences close to 1.0 SD . Effect sizes of this magnitude can be considered “very large” in psychology (Funder and Ozer (2019), raising suspicion when they emerge seemingly “randomly” in exploratory research on individual differences. Additionally, the fact that one cluster outperforms the other(s) in all variables simultaneously might also indicate an inflation in the number of detected latent classes (Toffalini et al., 2022). However, all correlations are quite strong (r in $[0.52, 0.70]$), and they are adequately modeled using a latent factor, so it is unlikely that they contribute to an inflation in the number of detected clusters. Instead, violations of assumptions may be problematic here. Three sum scores from mathematical tasks are used as indicators. One skewness coefficient is large (0.98 in math fluency), and another is moderately large (-0.68 in numerical operations). All kurtosis coefficients are moderately large (absolute values ≥ 0.50). It is worth noting that large skewness here likely reflects psychometric characteristics of the task and the use of sum scores. Note that math fluency has $M = 14.19$, $SD = 10.34$, thus the mean value is much closer to the lower (0) than to the upper (48) bound, and it is only 1.37 SD s away from the lower bound.

Data Simulation

As a retrospective evaluation of the inferential risks in the research scenario presented by Muñoz et al. (2023), Monte Carlo simulations were conducted. Data simulation was based on the coefficients

reported by the authors in their descriptive statistics table and was executed using the “semTools” package of R, which allows for the simulation of multivariate (non-)normal distributions. The sample size ($N = 428$), correlations, skewness, and kurtosis coefficients for the three measures of math ability used for clustering (i.e., math fluency, math problem solving, and numerical operations) were extracted from the original articles. Commendably, the authors provided the Mplus script that they used for fitting models, which could be reused in the present Monte Carlo simulation analyses. However, as Mplus is licensed, non-open-source software, making it difficult for others to freely replicate the results, additional simplified examples, entirely run with the freely available R software (R Core Team, 2010) are offered to further clarify the points, and the code is entirely provided.

Results

Ten datasets were simulated, and factor mixture models were fitted using the same Mplus code provided by Muñoz et al. (2023). Models featuring 1, 2, and 3 latent classes were fitted. Non-invariant model alternatives were preferred both to ensure maximum flexibility and because these were the models chosen by the authors in their final selected solution. In all 10 iterations, BIC consistently favored a three-class solution. The two-class model consistently outperformed the one-class model (median $\Delta\text{BIC} = -134.7$), and the three-class model consistently outperformed the two-class model (median $\Delta\text{BIC} = -28.26$). In all but one case, the likelihood ratio test also suggested that the one-class hypothesis (H_0) should be rejected in favor of a two-class solution ($ps < 0.001$), while in five of these cases, the two-class hypothesis (as H_0) was rejected in favor of the three-class solution ($ps < 0.05$). Results are presented in Table 1.

As a further sanity check, all 10 datasets were re-simulated with perfectly normal distributions (i.e., with all skewness and kurtosis coefficients set to zero in the data-generating process). In this second case, the BIC consistently (and correctly) favored the one-class solution (median $\Delta\text{BIC} = +17.94$ in favor of the one-class over the two-class solution, and median $\Delta\text{BIC} = +19.50$ in favor of the two-class over the three-class solution). Results are presented in Table 2. This double check confirmed that the inflation of the number of detected clusters was due to non-normality. Caution should be used, however, as several warnings emerged indicating that the latent variable covariance matrix was not positive definite in one or more classes, especially with the three-class solutions.

Table 1.

Results from 10 iterations of running factor mixture model, 1-3 classes, on simulated data with 3 indicators and parameters taken from Munez et al. (2023). Data present significant skewness and kurtosis (see text). There are no true latent classes, but the BIC consistently and correctly favors the 3-class solution. $N = 428$.

Iteration	BIC (1 class)	BIC (2 classes)	BIC (3 classes)	LRT p-val (2 vs 1 cl.)	LRT p-val (3 vs 2 cl.)
1	3260.71	3117.46	3058.55	<0.001	<0.001
2	3231.70	3120.85	3102.48	<0.001	0.002
3	3244.15	3101.92	3062.66	<0.001	<0.001
4	3123.03	2999.45	2979.78	<0.001	0.005
5	3189.74	3050.75	3022.55	<0.001	0.275
6	3164.93	3035.99	3011.98	<0.001	0.052
7	3228.02	3051.49	3048.14	<0.001	0.076
8	3131.17	2978.36	2931.72	<0.001	<0.001
9	3140.58	3024.68	2996.36	0.113	0.010
10	3199.68	3069.27	3016.92	<0.001	0.055

Table 2.

Results from 10 iterations of running factor mixture model, 1-3 classes, on simulated data with 3 indicators and correlations taken from Munez et al. (2023), but without any skewness and kurtosis (all parameters were fixed to zero). There are no true latent classes, and the BIC consistently favors the 1-class solution. $N = 428$.

Iteration	BIC (1 class)	BIC (2 classes)	BIC (3 classes)	LRT p-val (2 vs 1 cl.)	LRT p-val (3 vs 2 cl.)
1	3143.40	3170.59	3184.05	0.408	<0.001
2	3138.86	3144.23	3161.86	0.013	0.011
3	3219.51	3239.62	3266.04	0.603	<0.001
4	3242.65	3261.33	3285.14	0.379	0.256
5	3170.24	3188.30	3216.56	0.074	<0.001
6	3165.85	3180.07	3199.54	0.087	<0.001
7	3180.55	3203.14	3222.67	0.443	0.025
8	3194.90	3212.72	3241.66	0.059	0.373
9	3180.20	3187.64	3193.73	0.036	0.139
10	3184.38	3199.03	3217.02	0.007	0.093

Additional analyses

For further clarification, additional Monte Carlo simulations of Gaussian mixture models were conducted in R. The code is provided on GitHub at https://github.com/EnricoToffalini/commentary_mixture_skewness. The well-known “mclust” package (Scrucca, Fop, Murphy, & Raftery, 2016) was utilized. Tested solutions featured 1 to 3 components (latent classes/clusters), with BIC serving as the criterion for determining the optimal solution. Initially, 1,000 iterations were executed, simulating datasets each with $N = 428$, featuring skewness and kurtosis as reported above. Notably, in all cases, data were drawn from a single population, with no true clusters present. Despite this, the one-class solution was never favored by the BIC, despite being the correct one. In 41.4% of iterations, the two-class solution was favored, and in the remaining 58.6% of iterations, the three-class solution was favored. Subsequently, another 1,000 iterations were run. In this set, the same N and correlations across variables were simulated, but all skewness and kurtosis coefficients were set to zero. In this case, the one-class solution was (correctly) selected as optimal in 100% of iterations. This suggests that skewness and kurtosis are indeed the underlying factors contributing to the inflated number of latent classes detected.

Discussion

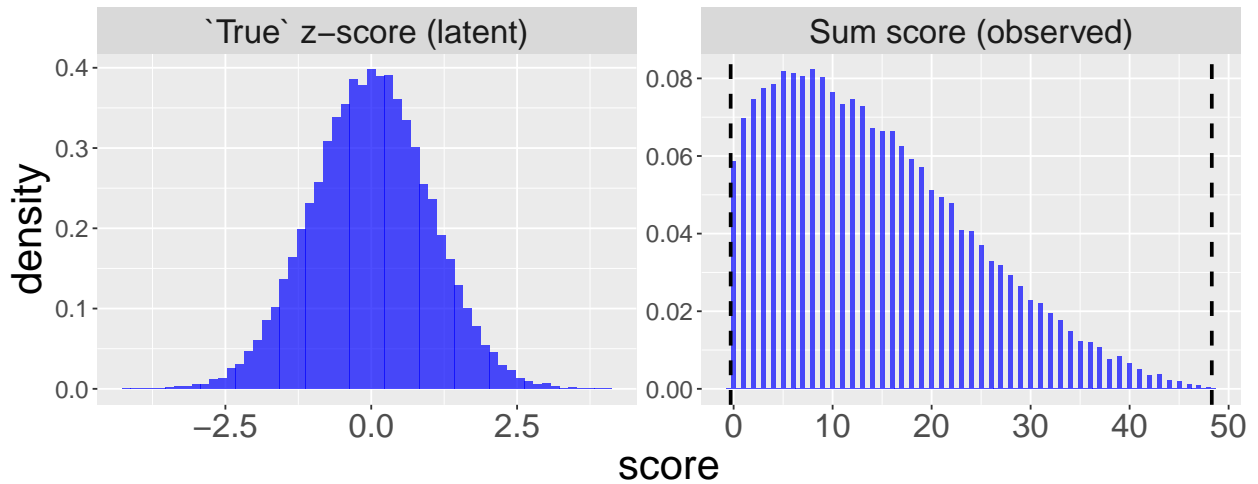
The results suggest that there may not be evidence of hidden heterogeneity in children with MLD. While the ground truth of the data-generating process is unknown concerning the real data, the present Monte Carlo simulation demonstrates that Muñoz et al. (2023) had a high chance of detecting multiple clusters even if no true latent classes existed within MLD. This article takes the opportunity to highlight a larger problem regarding clustering individuals based on psychological data. The main issue, in this case, appears to be the data not meeting the normality assumption. As explained earlier, non-normality may simply reflect the characteristics of the tasks used, and it is a widespread problem in psychometrics (e.g., Micceri, 1989). When assessing achievement or cognitive abilities in developmental samples, sum scores are often computed. As binomial distributions, sums of binomial (correct/incorrect) responses may lead to skewness unless averaging around 50%, and their mean and SD are non-independent (e.g., Edwards, 1960).

To clarify how sum scores from a binomial distribution may lead to skewness, consider Figure 1. On the left panel, the distribution of $N = 50,000$ simulated cases on a perfectly normal latent trait (“true” z-scores; $M = 0$, $SD = 1$, range in $[-Inf, +Inf]$) is shown. On the right panel, the

observed distribution of sum scores is displayed, assuming that individuals completed a 48-item task with binomial responses. Items were simulated to be moderately difficult on average ($M = 1$, $SD = 1$), resulting in right skewness of scores. The probit function was used to transform latent z-scores into probabilities. The observed sum scores present $M = 14.24$, $SD = 9.89$, and a skewness coefficient of -0.67. Note that these parameters are quite similar to those reported for math fluency in Muñez et al. (2023), albeit somewhat less skewed. The R code for plotting this figure and the underlying statistical computation is available online at: https://github.com/EnricoToffalini/commentary_mixture_skewness. To address this issue, researchers might preliminarily compute estimates of individual ability using Item Response Theory (IRT) models or via predicted scores with Confirmatory Factor Analysis (CFA) with ordinal items; these may be preferred solutions for normalizing scores.

Figure 1.

Example of distributions of true latent z-scores and corresponding observed sum scores in 50,000 simulated cases. Dashed vertical lines show the bounds of observed scores.



Moving forward, it is crucial to approach clustering techniques with caution in psychological research, especially when inference is the goal. This is particularly critical as clusters emerging via data-driven methods cannot easily be corroborated by other sources of evidence, and ground truth remains unknown. No descriptive statistics can back up inference unless visual inspection of scatter plots clearly suggests multimodal distributions. Unfortunately, this is unlikely to be the case in psychological research on individual differences, as it would require extremely large effect sizes for clusters to visually emerge from plotting alone. To clarify, distributions separated well more than

1 SD apart (i.e., Cohen's $ds \gg 1.0$) are required for observing multimodality in data distributions (e.g., ?, ?). When uncertain, a researcher may run Monte Carlo simulations as a good a priori sanity check to see if the chosen clustering methods provide valid results when the ground truth is known, considering the characteristics of the data. Example of R code has been provided for it. In addition to quantifying a “*type I* error of clustering”, that is the risk of detecting multiple classes when in fact only one exist, the readers may want to quantify statistical power, that is the probability of detecting the correct number of classes when true clusters/latent classes exist (Tein et al., 2013). As shown by simulation studies (e.g., Tein et al., 2013), however, sufficient power requires effect sizes of about Cohen's $d = 0.80$ or above in many independent dimensions simultaneously, which is bordering on credibility in psychology.

Code and data availability

Code and data used in this article are fully available on GitHub at: https://github.com/EnricoToffalini/commentary_mixture_skewness

References

- Astle, D. E., Holmes, J., Kievit, R., & Gathercole, S. E. (2022). Annual research review: The transdiagnostic revolution in neurodevelopmental disorders. *Journal of Child Psychology and Psychiatry*, 63(4), 397–417.
- Edwards, A. (1960). The meaning of binomial distribution. *Nature*, 186(4730), 1074–1074.
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in methods and practices in psychological science*, 2(2), 156–168.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, 105(1), 156.
- Muñoz, D., Bull, R., Lee, K., & Ruiz, C. (2023). Heterogeneity in children at risk of math learning difficulties. *Child Development*.
- R Core Team. (2010). R: A language and environment for statistical computing. (*No Title*).
- Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1), 289.
- Tein, J.-Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural equation modeling: a multidisciplinary journal*, 20(4), 640–657.
- Toffalini, E., Girardi, P., Giofrè, D., & Altoè, G. (2022). Entia non sunt multiplicanda... shall i look for clusters in my cognitive data? *Plos one*, 17(6), e0269584.