J. Intell. **2014**, *2*, 156-179; doi:10.3390/jintelligence2040156

OPEN ACCESS

Journal of
Intelligence
ISSN 2079-3200
http://www.mdpi.com/journal/jintelligence

Article

An Investigation of Growth Mixture Models for Studying the Flynn Effect

Grant B. Morgan * and A. Alexander Beaujean

Educational Psychology Department, Baylor University, One Bear Place #97301, Waco, TX, USA

* Author to whom correspondence should be addressed; E-Mail: Grant_Morgan@baylor.edu; Tel.: +1-254-710-7231

External editor: Joseph L. Rodgers

Received: 10 February 2014; in revised form: 11 September 2014 / Accepted: 24 September 2014 / Published: 16 October 2014

Abstract: The Flynn effect (FE) is the well-documented generational increase of mean IQ scores over time, but a methodological issue that has not received much attention in the FE literature is the heterogeneity in change patterns across time. Growth mixture models (GMMs) offer researchers a flexible latent variable framework for examining the potential heterogeneity of change patterns. The article presents: (1) a Monte Carlo investigation of the performance of the various measures of model fit for GMMs in data that resemble previous FE studies; and (2) an application of GMM to the National Intelligence Tests. The Monte Carlo study supported the use of the Bayesian information criterion (BIC) and consistent Akaike information criterion (CAIC) for model selection. The GMM application study resulted in the identification of two classes of participants that had unique change patterns across three time periods. Our studies show that GMMs, when applied carefully, are likely to identify homogeneous subpopulations in FE studies, which may aid in further understanding of the FE.

Keywords: Flynn effect; growth mixture model; National Intelligence Tests; Estonia

---

1. Introduction

The Flynn effect (FE) is the well-documented generational increase of mean IQ scores found in many countries [1,2]. The FE was observed as early as the 1930s, but only began to be systematically studied in the 1980s [3]. While the typical FE is between three to five IQ points per decade, the effect's magnitude and direction have shown considerable variation across time and location [4].

Throughout its history, the FE has been fraught with methodological issues. For example, many of the original FE studies were done comparing mean scores without paying attention to the scores' distributions [5]. Another criticism is that the compared scores were often from instruments normed at different time points, with no investigation into whether the scores were actually comparable [6,7]. While some current FE investigations have sought to use more robust methods, e.g., [8–10], there are still some methodological issues that remain.

One methodological issue that has not received much attention in the FE literature is heterogeneity in growth patterns. Cattell [11] first discussed this over 60 years ago, but methods for assessing differential change have substantially evolved since his writing. While subsequent studies have reported different magnitudes and directions of the FE (e.g., [4,12,13]), almost all of them assume homogenous change within the given sample (for an exception, see [14]). Thus, these studies do not investigate whether there are any differences within the FE magnitude or direction. One way to investigate whether there are differences in the FE within a given sample is to use growth mixture models.

1.1. Mixture Models

Mixture modeling is a model-based approach used to identify underlying classes (also called subgroups, profiles or populations) within a dataset that are unknown *a priori* [15]. That is, they are typically used to explore unobserved heterogeneity when it is theorized to exist or when observed (*i.e.*, known) groups are not of primary interest in the analysis. When mixture models are used with latent variables, they are similar to multigroup models, with the major difference being that the group membership is not directly observed with mixture models. Instead, mixture models allow group membership to be a latent variable, measured indirectly from the observed variables [16].

Mixture models can be used with data collected at one time point (*i.e.*, cross-sectional) or multiple time points (*i.e.*, longitudinal data). Mixture models based on longitudinal data are generally referred to as growth mixture models. Growth mixture models (GMMs) combine latent curve models [17] and mixture models. Thus, the objective of fitting GMMs is twofold [18]. First, GMMs describe, in a *post-hoc* manner, possible classes within the data. This is similar to exploratory factor analysis in that the number of classes are unknown at the outset and must be subsequently determined by the data. Second, once the number of classes are found, then GMMs examine how the indicator variables change over time, as well as the differences in this change between and within the unobserved classes.

As with other latent variable models, GMMs have the benefit of estimation flexibility. Specifically, researchers have the option to freely estimate or constrain any of the model parameters based on the theory of how the variables should relate [19], although the complexity of GMMs does somewhat limit this flexibility [20]. Another benefit of GMMs is the ability to compare the fit of competing models to the data [21,22]. Model fit is especially important in GMMs, as it can be used to guide the number of

classes to retain [23]. There are many fit indices available to examine fit in a GMM, each using different criteria to examine the fit of the model to the data. We discuss these different fit indices in Section 2.3.

1.2. Current Study

The purpose of the current study is to investigate the performance of the various measures of model fit for GMMs in data that resemble previous FE studies. Using Monte Carlo methods [24,25], we systematically examined the utility of fit indices to identify the correct number of underlying classes while varying the sample size, score reliability, proportions of participants in each class (*i.e.*, class prevalence) and growth pattern. Based on the results of the Monte Carlo study, we then fit a GMM to data from the National Intelligence Tests [26].

2. Method

2.1. Monte Carlo Study

Monte Carlo (MC) studies are experiments that repeatedly generate sample data from a population model [24,25]. The purpose of the repeated sample generation is to obtain empirical sampling distributions [27]. MC methods are commonly used in latent variable modeling for multiple reasons, one of which is to examine the performance of specific modeling procedures under a particular set of known circumstances [28]. Use of a population with a known structure allows for the investigation of fit index performance. To that end, we examined the utility of GMM fit indices under specific conditions that have been, or are likely to be, observed in investigations of the FE.

2.1.1. Population Models

The population models from which samples were drawn for the current study all contained two classes of respondents defined by their growth patterns across three time points. We used two classes for the population models, because it is the simplest mixture model scenario, which is best placed to start when applying new models. Extensions to situations with more than two classes are easily generalizable from our presentation and have been documented elsewhere (e.g., [23,29,30]). We used three time points because that is the minimum needed for a latent model [17]. In one set of population models, both classes exhibited linear trends; in the other set of population models, both classes exhibited quadratic trends.

A typical growth model consists of two parts: a Level-1 component and Level-2 component [31]. The Level-1 component is comprised of the observed variable (e.g., intelligence test scores) measured at the different time periods. The Level-1 portion of the growth model is given in Equation 1.

$$Y_{ij} = \lambda_{0i} + \lambda_{1i}(T_j) + \lambda_{2i}(T_j^2) + \epsilon_{ij}, \quad (1)$$

where Y_{ij} is the value of the outcome variable for person i at time period j, λ_{0i} is the true intercept of the change trajectory for person *i*, λ_{1i} is the true linear slope of the change trajectory, λ_{2i} is the true quadratic curvature of the change trajectory, ϵ_{ij} is the residual for person i and T_j is a variable indicating the time period.

The GMM version of the growth model is very similar to Equation 1, but allows the model to differ by class. It is shown in Equation 2.

$$Y_{ijk} = \lambda_{0ik} + \lambda_{1ik}(T_j) + \lambda_{2ik}(T_j^2) + \epsilon_{ik}, \tag{2}$$

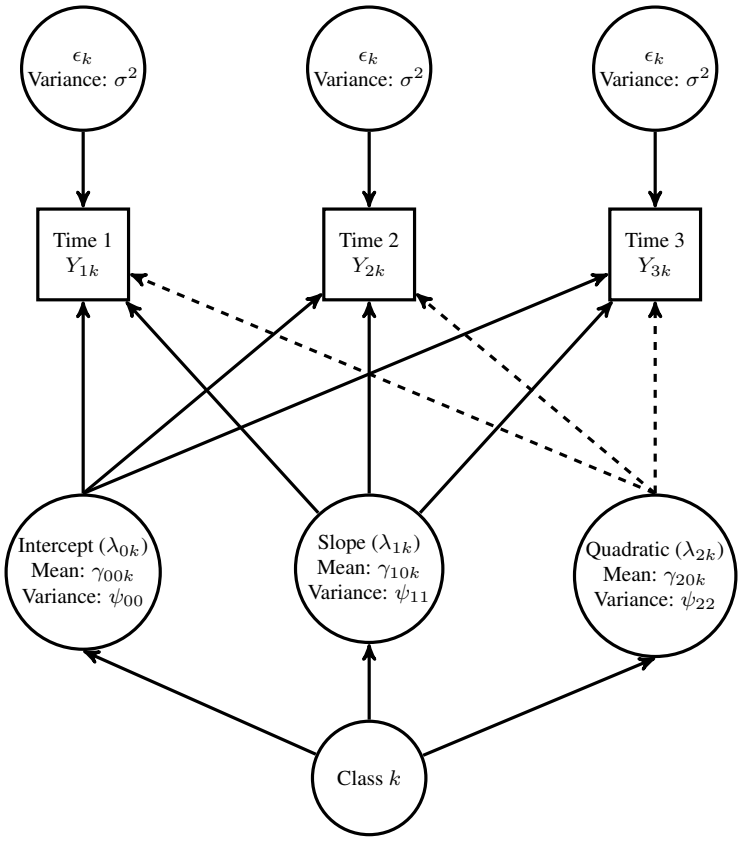
where k represents the class to which person i belongs.

The variables used in Equation 1 are not independent of each other, since they are observed on the same individuals at each time period. Thus, the Level-2 components are the unique individuals in the sample. The Level-2 model for the k classes is given in Equation 3.

$$\begin{aligned} \lambda_{0ik} &= \gamma_{00k} + \zeta_{0ik} \\ \lambda_{1ik} &= \gamma_{10k} + \zeta_{1ik} \\ \lambda_{2ik} &= \gamma_{20k} + \zeta_{2ik}, \end{aligned} \tag{3}$$

where γ_{00k} is the mean intercept for class k, ζ_{0ik} is the intercept random error for person i in class k (with variance of ψ_{00}), γ_{10k} is the mean of the linear slopes for class k, ζ_{1ik} is the linear slope random error for person i in class k (with variance of ψ_{11}), γ_{20k} is the mean of the quadratic terms for class k and ζ_{2ik} is the quadratic slope random error for person i in class k (with a variance of ψ_{22}). For models that did not include the quadratic term, γ_{20k} was set to zero, and the quadratic slope variance (ζ_{2ik}) was removed. Figure 1 shows a conceptual diagram of the GMM. For all models in the MC study, no linear or quadratic slope variances were estimated.

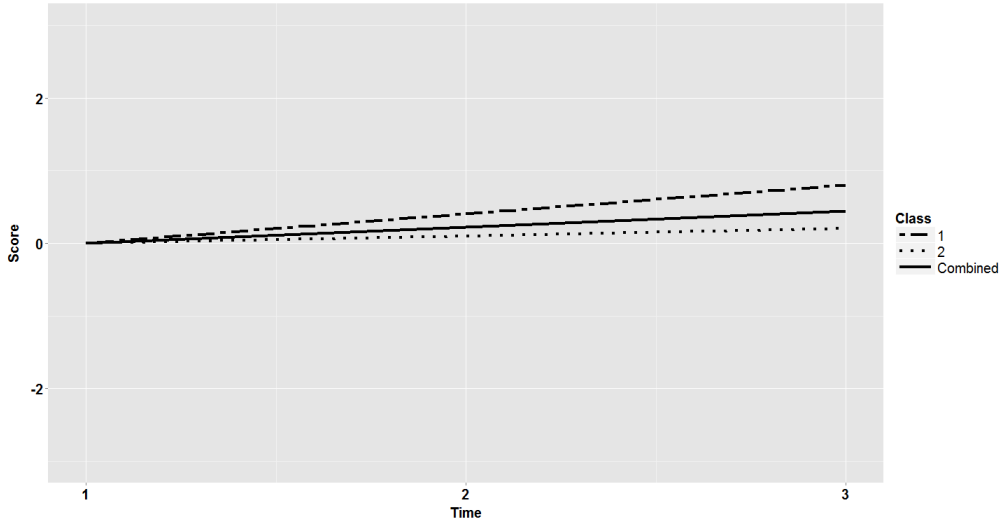
Figure 1. Conceptual diagram of the linear and quadratic growth mixture model with Level-1 subscripts removed. Coefficients for dashed lines were set to zero for the linear model.



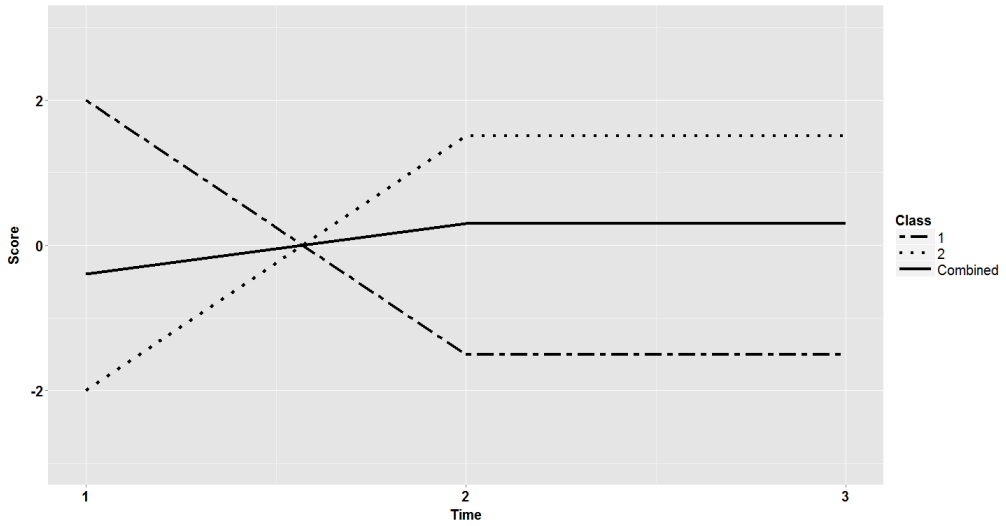
In all population models, the class trends were specified such that if classes were ignored, the change in scores across time resembled a typical FE (*i.e.*, a rise of approximately 0.30 IQ points a year). For example, the population version of the class-specific growth models are depicted in Figure 2 along with the combined growth model if classes were ignored. Figure 2a represents the model that included the linear term, while Figure 2b represents the model that included the linear and quadratic terms. We refer to the conditions with the quadratic growth term as the crossing growth pattern conditions, while we refer to the conditions with only the linear growth term as the non-crossing growth pattern conditions.

We chose the two growth patterns, because they represented two extremes for having heterogeneity in the FE, which is a recommended strategy when initially investigating a phenomenon [32]. Our choice does not indicate that we believe they are the only two growth patterns, only that they are two possible forms of heterogeneous growth that could underlie the FE.

Figure 2. Class-specific and combined growth models. (**a**) Linear (non-crossing) growth model; (**b**) quadratic (crossing) growth model.



(a)



(b)

2.1.2. Design Factors

Table 1. Design factors for the Monte Carlo study.

Factor	Level 1	Level 2
Class Prevalence (Class 1/Class 2)	0.40/0.60	0.30/0.70
Sample Size	200	800
Measure Reliability	0.80	0.95
Growth Pattern	Linear (Not Crossed)	Quadratic (Crossed)

Note: For all models, the linear slope and quadratic term variances were set to zero, the covariances were set to zero, the intercept variances were set to 0.80 and the residual variances were held constant across all time periods.

The manipulated design factors included in the MC study were: (a) class prevalence, (b) sample size, (c) score reliability and (d) growth pattern. The levels for each factor are given in Table 1. The design was fully crossed to yield 16 cells (*i.e.*, unique conditions). In all models, the intercept variance was set to 0.80. The reliability was specified as the ratio of the observed variable's true variance (*i.e.*, variance explained by the model) to the total variance (*i.e.*, true variance plus residual variance) of variables at each time period. The reliability formula is given in Equation 4.

$$\rho_{kt} = \frac{\psi_{00k} + \lambda_t^2\psi_{11k} + \lambda_t^4\psi_{22k} + 2\lambda_t\psi_{01k} + 2\lambda_t^2\psi_{02k} + 2\lambda_t^3\psi_{12k}}{\psi_{00k} + \lambda_t^2\psi_{11k} + \lambda_t^4\psi_{22k} + 2\lambda_t\psi_{01k} + 2\lambda_t^2\psi_{02k} + 2\lambda_t^3\psi_{12k} + \sigma_{kt}^2}, \tag{4}$$

where ψ_{00} is the intercept variance in class k, ψ_{11} is the linear slope variance in class k, ψ_{22} is the quadratic term variance in class k, ψ_{01} is the covariance between the intercept and linear slope in class k, ψ_{02} is the covariance between the intercept and quadratic term in class k, ψ_{12} is the covariance between the linear slope and quadratic term in class k, λ_t is the time period indicator for time t and $\sigma_k^2 t$ is the residual variance at time t in class k.

For the current study, the linear slope and quadratic term variances were set to zero in both classes, as were the covariances. Furthermore, the intercept variances were set to 0.80 in both classes, and the residual variances were held constant across all time periods. Therefore, Equation 4 reduces to Equation 4a.

$$\rho = \frac{\psi_{00}}{\psi_{00} + \sigma^2} \tag{4a}$$

As an example, for a reliability of 0.95 for the linear (non-crossing) model, plug in the known values into Equation 4a and solve for the unknown value, σ^2:

$$0.95 = \frac{0.80}{0.80 + \sigma^2} \Rightarrow \sigma^2 = 0.0421\,.$$

The population effect size for the GMMs is given in Equation 5.

$$d = \frac{\mu_D}{\sigma_D} = \frac{\left(\sum_{k=1}^{K}\pi_k\mu_{3k}\right) - \left(\sum_{k=1}^{K}\pi_k\mu_{1k}\right)}{\sqrt{\left(\sum_{k=1}^{K}\pi_k(\mu_{3k} - \mu_D)^2\right) + 2\sigma^2}}, \tag{5}$$

where μ_D is the mean of the difference scores, σ_D is the standard deviation of the difference scores, π_k is the prevalence of class k, μ_{3k} is the mean of class k at Time Period 3, μ_{1k} is the mean of class k at Time Period 1 and σ_t^2 is the residual variance at time period t.

As an example, the effect size for the linear (non-crossing) model with a reliability of 0.80 and a Class 2 prevalence and slope of 0.60 and 0.10, respectively, and a Class 1 prevalence and slope of 0.40 and 0.40, respectively is:

$$d = \frac{\mu_D}{\sigma_D} = \frac{(\pi_2\mu_{13} + \pi_1\mu_{23}) - (\pi_2\mu_{11} + \pi_1\mu_{21})}{\sqrt{\pi_2(\mu_{13} - \mu_D)^2 + \pi_1(\mu_{23} - \mu_D)^2 + 2\sigma^2}} =$$

$$\frac{(0.60 * 0.20 + 0.40 * 0.80) - (0.60 * 0 + 0.40 * 0)}{\sqrt{0.60(0.20 - 0.44)^2 + 0.40(0.80 - 0.44)^2 + (2 * 0.20)}} = 0.63\,.$$

The population effect size for each cell of the design is given in Table 2.

Table 2. Population effect size for each cell of the design.

			Growth Pattern	
Class Prevalence	**Reliability**	**Sample Size**	**Crossing**	**Non-Crossing**
π_1 = 0.40, π_2 = 0.60	0.80	200	0.20	0.63
π_1 = 0.40, π_2 = 0.60	0.80	800	0.20	0.63
π_1 = 0.40, π_2 = 0.60	0.95	200	0.20	1.07
π_1 = 0.40, π_2 = 0.60	0.95	800	0.20	1.07
π_1 = 0.30, π_2 = 0.70	0.80	200	0.43	0.55
π_1 = 0.30, π_2 = 0.70	0.80	800	0.43	0.55
π_1 = 0.30, π_2 = 0.70	0.95	200	0.43	0.95
π_1 = 0.30, π_2 = 0.70	0.95	800	0.43	0.95

Note: Effect sizes are given in the last two columns.

2.2. Data Generation

The population structure used in the current study is a mixture model with known numbers of classes, class prevalence, the class-specific growth model and reliability (see Table 1). Samples were generated from these structures, and GMMs were then fit to the sample data to determine the extent to which the true structure was recovered. For each cell of the design, 1,000 samples were simulated using five hundred random starting values. Use of a high number of random starting values for model estimation improves the chances that a researcher identifies the global maximum of the likelihood function [33].

2.3. Fit Indices to Aid in the Growth Mixture Model Selection

The fit indices for GMMs reflect either absolute model fit, relative fit or classification certainty [34].

2.3.1. Absolute Model Fit

Absolute fit indices illustrate how well the model with a given number of classes fits the data irrespective of how other models fit the data. A common absolute fit index is the minimized fit function value (*i.e.*, likelihood). The logarithm of the likelihood (log-likelihood) is often used rather than the likelihood index itself for greater simplicity in interpretation. Larger values of the log-likelihood value (*i.e.*, values closer to zero, because they are negative) provide stronger evidence that the model fits the data. In GMMs, comparing log-likelihood values of competing models is likely to prove unfruitful, because the log-likelihood may prefer models with more underlying classes. Models with more underlying classes violate the parsimony principle and would also likely be difficult to interpret [34]. The log-likelihood is still important, because it is used as part of other fit indices, such as information criteria, which are discussed in Section 2.3.3.

2.3.2. Lo–Mendell–Rubin Likelihood Ratio Test

Traditionally, comparing nested models via the likelihood ratio test (LRT) is common practice in latent variable modeling [35]. This LRT cannot be conducted in the same manner with GMMs, because the likelihood ratio difference between models with different numbers of classes typically does not follow a χ^2 distribution [33]. Furthermore, the traditional LRT of nested models examines differences between two models that only differ in the model parameterization. In GMMs, competing models may specify different numbers of underlying classes. Models that differ in the number of classes necessarily have different model parameterizations, which makes the traditional test of nested models inappropriate.

Lo, Mendell and Rubin [36] extended the previous work of Vuong [37], who derived a likelihood ratio test for model selection. The eponymous Lo–Mendell–Rubin likelihood ratio test (LMR) is a global test of model fit in which the LRT distribution is approximated, thus making the comparison of neighboring class models possible [36]. They developed an *ad hoc* adjustment for finite sample sizes that improves the accuracy of inferences based on the test. The LMR compares the improvement in fit between neighboring class models (*i.e.*, compares a model with $k-1$ classes against one with k classes). The resulting statistic's p-value can be used to test if the increase in model fit from specifying an addition class is unlikely to be due to chance alone [29]. The best fitting model according to the LMR is the last model that resulted in the rejection of the null hypothesis.

2.3.3. Relative Fit Indices

Relative fit indices are interpreted in relation to the fit estimates from other models. These indices are computed by imposing a penalty on the log-likelihood function when more parameters are estimated and/or when fewer subjects are included in the analysis.

Most relative fit indices used to assess a GMM are information-theoretic indices [38]. These indices are related to model likelihood in that minimizing the information indices coincides with maximizing likelihood. Unlike likelihood-based procedures, the information-theoretic indices can be applied to non-nested models [39,40]. We list these indices in Table 3.

Table 3. Information-theoretic fit indices.

Index	Abbreviation	Formula
Akaike information criteria	AIC	$-2\log L + 2p$
Consistent AIC	CAIC	$-2\log L + p[\log(n) + 1]$
Corrected AIC	AICc	AIC $+\frac{(2(p+1)(p+2))}{(n-p-2)}$
Bayesian information criteria	BIC	$-2\log L + p\log(n)$
Sample size adjusted BIC	SSBIC	$-2\log L + p\log\left[\frac{(n+2)}{24}\right]$
Draper's information criterion	DIC	$-2\log L + p\left(\log\left(\frac{n}{2\pi}\right)\right)$

Note: L, likelihood; p, number of free (estimated) parameters; n, sample size.

One of the original information-theoretic indices was the Akaike information criteria (AIC), which adds a complexity penalty to a model's log-likelihood value [41]. The consistent AIC (CAIC; [42]), corrected AIC (AICc; [43]), Bayesian information criteria (BIC; [44]), sample size adjusted BIC (SSBIC; [45]) and Draper's [1995] information criterion (DIC) are all similar to the AIC, but the penalties they impose for complexity are all functions of the sample size. For all information-based criteria, comparatively lower values indicate better fitting models [47], although all criteria may not be minimized for the same model [34].

Prior simulation studies have shown that AIC tends to overestimate the correct number of underlying mixtures [29,48–52], despite being a frequently reported index. The BIC provides information for both selecting models with varying numbers of mixtures and selecting between competing models that are parameterized differently, such as restrictions made to the covariance matrices [33]. The BIC tends not to overestimate the number of underlying mixtures as frequently as the AIC does, because it imposes a stronger complexity penalty. Under smaller sample size conditions, Celeux and Soromenho [48] and Tofighi and Enders [23] found that BIC tended to underestimate the number of classes using sample sizes of 300 and 400, respectively. Morgan [50] found that the BIC and SSBIC tended to identify a higher proportion of correct solutions, except under conditions with very rare classes. Yet, when rare classes were present, none of the indices examined performed well [50]. Similar to Celeux and Soromenho [48] and Tofighi and Enders [23], Morgan [50] found that when BIC identifies the incorrect mixture model, it underestimates the number of mixtures. Nylund *et al.* [29] found that BIC generally outperformed all other information-based criteria, including AIC, CAIC and SSBIC. Yang [52] found SSBIC to outperform AIC, CAIC and BIC in models with categorical indicators and at least 50 subjects per latent class. These findings suggest that SSBIC may be related to class enumeration and sample size in order to maintain a certain subject-to-class ratio. The AICc and DIC have not been as systematically studied as the other criteria, at least when used with GMMs.

2.3.4. Classification Certainty

Assigning individuals to the most appropriate class is an important component of GMMs. If the classes are distinct and well defined, then all observations will only have one class to which they belong. When the respondents are likely to belong to a number of classes, then the solution lacks meaningful classification. Consequently, GMM solutions that classify the respondents unambiguously represent

better models. The overall distinctiveness of the GMM's classification is commonly expressed with entropy-based measures [53,54].

Entropy is defined as:

$$O_k = -\sum_{i=1}^{n}\sum_{k=1}^{K} \pi_{(k|y_i)} \log(\pi_{(k|y_i)}), \tag{6}$$

where $\pi_{(k|y_i)}$ is the estimated probability that person i belongs to class k, K is the number of classes and n is the sample size. Values of O_k closer to zero indicate better classification.

Relative entropy [55], E_k, is a scaled version of O_k, defined in Equation 7.

$$E_k = 1 - \frac{O_k}{n \log(K)}, \tag{7}$$

Unlike O_k, E_k is bounded by zero and one, with larger values indicating better classification [56].

Some disadvantages of using entropy (or relative entropy) are that it assumes that the estimated model is correct and there is no agreed level for acceptable values [57]. Another disadvantage is that it can be negatively influenced by chance misclassification for models with a higher number of classes [34]. That is, there is a greater opportunity for misclassification for models with more classes, which may result in the entropy-based estimates spuriously indicating that a model fits the data poorly.

The integrated classification likelihood criteria [58] is another classification-based fit index. Its calculation is complex, so it is often estimated using a BIC-type approximation (ICL-BIC). When class sizes are sufficiently large, the performance of the ICL-BIC differs very little from the more complex version [58]. The ICL-BIC is defined in Equation 8.

$$\text{ICL-BIC} = -2 \log L + 2O_k + p \log(n), \tag{8}$$

where p is equal to the number of free parameters, O_k is entropy, and n is the sample size. While the ICL-BIC has not been widely examined, studies that have examined it have found that it works well, often outperforming other fit indices [59,60].

2.4. Analysis

We fit two through five class solutions to each generated dataset. There were no replications that produced implausible values or that did not converge. Descriptive information (e.g., proportions of correct model selection) was adjusted accordingly based upon the number of usable replications for a given cell.

For each usable replication, we recorded the AIC, AICc, CAIC, BIC, SSBIC, DIC, ICL-BIC, relative entropy and LMR fit measures and used them to identify the best model. For AIC and its variants, as well as the ICL-BIC, the model with the smallest values for each index, was selected as the best fitting model. For each LMR test, we started by testing the null hypothesis that the one-class model fit the data better than the two-class model. If this null hypothesis was rejected (using a Type I error rate of 0.05), then we tested the null hypothesis that the two-class model fit the data better than the three-class model.

We repeated this process, until we either failed to reject the null hypothesis or reached five classes. We computed the mean and quartiles of the entropy values for the two-class models.

Next, the model identified as the best was compared against the population model. We assigned correctly flagged solutions a one and incorrectly flagged solutions a zero. Therefore, for each condition and fit index, the numbers of correct and incorrect solutions were recorded. The proportion of correctly identified solutions was reported for each fit index, so higher mean values indicate higher rates of accuracy.

To identify design factors that had an effect on the correct identification of the number of classes, we entered the design factors and their two-way interactions into logistic regression models predicting the accuracy of each fit index. The odds ratio (OR) from the logistic regression output were used as estimates of the effect size of each design factor. Finally, since the LMR uses a hypothesis test, this allows for an investigation of the test's Type I error rate and statistical power.

2.5. Software

We used *Mplus* [61] to generate the data and estimate the parameters via the MplusAutomation package [62] in R [63]. Parameters were estimated using maximum likelihood estimation with robust standard errors [61].

3. Results

3.1. Monte Carlo Study

A total of 16,000 datasets were successfully generated (16 conditions × 1,000 replications), and the general structure of the samples closely approximated the structure specified by the population model. In nearly every dataset, the structure of the sample matched that specified by the population model within two-thousandths of the true parameters, on average. Convergence rates were 100% for all models. The accuracy for each fit index across all cells is presented in Table 4, while the accuracy within the crossing and non-crossing pattern conditions is presented in the top and bottom parts of Table 5, respectively.

Table 4. Overall model selection accuracy rates across all conditions.

Fit Index	Accuracy (% Correct)
Akaike information criterion (AIC)	57.1
Corrected Akaike information criterion (AICc)	66.7
Consistent Akaike information criterion (CAIC)	99.9
Bayesian information criterion (BIC)	99.8
Sample size adjusted BIC (SSBIC)	79.0
Draper's information criterion (DIC)	96.8
Integrated classification likelihood with BIC approximation (ICL-BIC)	89.5
Lo–Mendell–Rubin likelihood ratio test (LMR)	54.4

Table 5. Accuracy (% correct) of all fit indices for crossing growth pattern conditions.

Class Prevalence	Reliability	Sample Size	Fit Index							
			AIC	AICc	CAIC	BIC	SSBIC	DIC	ICL-BIC	LMR
Crossing Growth Pattern										
$\pi_1 = 0.40, \pi_2 = 0.60$	0.80	200	60	75	100	100	66	96	100	76
$\pi_1 = 0.40, \pi_2 = 0.60$	0.80	800	58	63	100	100	96	100	100	73
$\pi_1 = 0.40, \pi_2 = 0.60$	0.95	200	60	76	100	100	67	96	100	78
$\pi_1 = 0.40, \pi_2 = 0.60$	0.95	800	60	64	100	100	96	100	100	73
$\pi_1 = 0.30, \pi_2 = 0.70$	0.80	200	58	74	100	100	64	96	100	78
$\pi_1 = 0.30, \pi_2 = 0.70$	0.80	800	58	62	100	100	97	100	100	74
$\pi_1 = 0.30, \pi_2 = 0.70$	0.95	200	58	74	100	100	66	96	100	79
$\pi_1 = 0.30, \pi_2 = 0.70$	0.95	800	60	63	100	100	97	100	100	73
Non-Crossing Growth Pattern										
$\pi_1 = 0.40, \pi_2 = 0.60$	0.80	200	50	66	100	99	58	92	66	9
$\pi_1 = 0.40, \pi_2 = 0.60$	0.80	800	59	63	100	100	96	99	73	13
$\pi_1 = 0.40, \pi_2 = 0.60$	0.95	200	55	68	100	100	60	93	77	24
$\pi_1 = 0.40, \pi_2 = 0.60$	0.95	800	62	65	100	100	96	99	94	80
$\pi_1 = 0.30, \pi_2 = 0.70$	0.80	200	50	66	100	100	57	93	68	9
$\pi_1 = 0.30, \pi_2 = 0.70$	0.80	800	57	61	100	100	96	99	74	17
$\pi_1 = 0.30, \pi_2 = 0.70$	0.95	200	50	65	100	99	58	93	85	29
$\pi_1 = 0.30, \pi_2 = 0.70$	0.95	800	60	64	100	100	94	99	97	84

Note: Values are rounded to the nearest percentage. AIC, Akaike information criterion; AICc, corrected Akaike information criterion; CAIC, consistent Akaike information criterion; BIC, Bayesian information criterion; SSBIC, sample size adjusted Bayesian information criterion; DIC, Draper's information criterion; ICL-BIC, integrated classification likelihood with Bayesian-type approximation; LMR, Lo–Mendell–Rubin likelihood ratio test.

3.1.1. Relative Fit Indices

The percentage of correctly identified solutions based on AIC across all design factors was 57.1%. The design factors with the strongest impact of AIC's accuracy were the class-specific growth patterns ($OR = 0.70$) and the interaction between the class-specific growth patterns and sample size ($OR = 1.42$). That is, when controlling for the other design factors, AIC tended to identify the correct model less frequently with the non-crossing growth pattern. Yet, the effect of the growth patterns was different, depending on the sample size. The cell of the design with the lowest AIC accuracy rate (49.7%) contained 200 observations, a reliability of 0.80, a class prevalence of 0.60 and 0.40 and non-crossing class-specific growth patterns. The cell of the design with the highest AIC accuracy rate (61.7%) contained 800 observations, a reliability of 0.95, a class prevalence of 0.60 and 0.40 and non-crossing class-specific growth patterns.

The effect the design factors had on the performance of the AICc was similar to that of AIC with the addition of the main effect of sample size also having a stronger impact ($OR = 0.56$). After controlling for the other design factors, AICc tended to perform better with the smaller sample size (n = 200) than the larger sample size (n = 800).

The CAIC and BIC were highly accurate at identifying the correct number of classes across all design factors.

While the SSBIC generally performed well, it was not as accurate as the CAIC or BIC. Sample size ($OR = 15.10$) and growth pattern ($OR = 0.76$) were the design factors that had the largest main effects. More specifically, after controlling for the other design factors in the model, the odds that SSBIC identified the correct class solution with a sample size of 800 were 15.1-times the odds that SSBIC identified the correct class solution with a sample of 200. Furthermore, the SSBIC was less accurate at identifying the correct number of classes in the non-crossing conditions than it was in the crossing conditions.

The DIC performance was very similar to the BIC. Unlike the BIC, though, DIC's performance was strongly influenced by two of the design factors: sample size ($OR = 13.00$) and growth patterns ($OR = 0.49$). Although DIC performed very well in the large sample size condition (n = 800; 99.4%), it performed well in the small sample size condition, too (n = 200; 94.2%).

3.1.2. Classification Certainty

Relative entropy was sensitive to the pattern of growth in each class. In all crossing growth pattern cells, the estimated mean relative entropy in the models containing two classes was 1.0, indicating perfect classification. In the non-crossing growth pattern cells, the estimated mean relative entropy in the two-class models was 0.60, and the quartiles of the relative entropy values were 0.50 (Q1), 0.58 (Q2) and 0.66 (Q3), with a minimum value of 0.20 and a maximum value of 1.0. Thus, relative entropy estimates were much higher with crossing growth patterns, which indicates that they were more distinguishable than the non-crossing growth patterns.

The overall accuracy of ICL-BIC was 89.5%, but this was influenced by the growth patterns and the reliability of design factors. For the crossing growth patterns, ICL-BIC reached 100% accuracy in all cells. Among the non-crossing growth pattern cells, ICL-BIC was generally more accurate when

reliability was higher (87.9%, OR = 9.37) and less accurate when reliability was lower (70.0%). ICL-BIC was least accurate (65.6%) in the cell with a smaller size (n = 200), lower reliability (0.80) and a growth pattern prevalence of 0.60 and 0.40.

3.1.3. Lo–Mendell–Rubin Likelihood Ratio Test

The LMR was sensitive to the growth patterns, reliability and sample size design factors. After controlling for the other design factors, the LMR tended to identify the correct number of classes more frequently: (a) in the larger sample size condition rather than the smaller sample size condition; (b) in the higher reliability condition rather than the lower reliability condition; and (c) in the crossing growth pattern condition rather than the non-crossing growth pattern condition.

To examine the Type I error rate of the LMR, we calculated the proportion of times the *p*-value was less than the pre-established Type I error rate of 0.05. Since the null hypothesis for the LMR is that the number of underlying classes is one less than the number being fit in a particular model, we examined the *p*-values for the model with three classes. The observed Type I error rate in this study was 0.16, which exceeded the maximum allowable rate. The inflation was primarily driven by the crossing growth pattern conditions in which the observed Type I error rate was 0.25. The observed Type I error rate in the non-crossing growth pattern conditions was 0.07, which was closer to the nominal Type I error rate used in the study.

To estimate the power of the LMR tests, we calculated the proportion of times we rejected the null hypothesis for the two-class solution. The observed power across all cells of the design was 0.68. The observed power in the crossing and non-crossing growth pattern conditions was 1.00 and 0.37, respectively. The observed power in the smaller and larger sample size conditions was 0.60 and 0.77, respectively. The observed power in the lower and higher reliability conditions was 0.56 and 0.80, respectively.

3.2. National Intelligence Tests

To demonstrate the use of growth mixture modeling for examining the FE, we fitted a series of GMMs to data collected in the National Intelligence Tests (NIT; [26]), which is a 10-sub-test intelligence measure modeled after the Army Alpha and Beta tests [64]. This data was collected in Estonia at three separate time points: 1934, 1998 and 2006. More information about the sample and the NIT can be found in previous studies using these data [7,9,65,66].

The IQ score modeled in the analysis was the summed score across the ten sub-tests. The observed reliability (α) coefficients of the summed score were 0.90, 0.84 and 0.86 for the year 1934, 1998 and 2006, respectively; very similar to the reliability estimates used in the MC study.

In the absence of longitudinal data from the exact same participants (*i.e.*, panel study), we used samples from the same population across three time periods (*i.e.*, trend study) [67]. To aid in the comparability of the participants, we matched participants from the three waves on the basis of age and sex. These were the only two demographic variables available in the dataset across all three waves that are related to cognitive ability. Thus, matching them is one way to statistically control for between-wave differences in these variables.

The total number of matched cases used in the applied analysis was 361, which is similar to the small sample size generated in the MC study. The mean and standard deviation of the IQ scores across the three waves of data are given in Table 6 and support the presence of an FE in the combined sample.

We fitted a series of two-to-five-class GMMs to the NIT IQ data just as we did in the MC study. To aid in model selection, we used the fit indices and criteria specified in Section 2.3.

The values of fit indices for the fitted models are presented in Table 7. The majority of the relative fit indices (BIC, CAIC, DIC and ICL-BIC) indicated that a two-class model fit the best; the AIC and SSBIC indicated that a four-class model fit the best, and the AICc indicated that a five-class model was the best. The LMR identified a one-class solution as the best fitting model. The MC study showed that the BIC, CAIC and DIC were the most accurate fit indices, and they all indicated the two-class solution as the best for the NIT data. Therefore, based on the fit index performance, we selected the two-class solution for interpretation.

Table 6. Descriptive statistics for latent classes.

Variable	Class 1	Class 2	Combined
	Membership		
n	89 (24.7%)	272 (75.4%)	361
Female	43 (48.3%)	169 (62.1%)	
Age			
12 Years	12 (13.5%)	35 (12.9%)	
13 Years	66 (74.2%)	155 (57.0%)	
14 Years	11 (12.4%)	82 (30.2%)	
	NIT Scores		
Time 1	236.0 (49.4)	238.2 (46.7)	237.6 (47.3)
Time 2	266.4 (36.0)	264.4 (36.1)	264.9 (36.0)
Time 3	223.0 (19.4)	287.9 (24.1)	271.9 (36.2)

Note: Membership values are sample sizes (% of total population). NIT (National Intelligence Test) scores are means (standard deviations).

The growth patterns of the two classes, as well as the growth pattern of the combined class, are shown in Figure 3. Descriptive statistics for the classes are given in Table 6. Class 1 was characterized by an initial positive growth pattern followed by a negative change from the second to the third time points, while Class 2 was characterized by positive growth across all three time points. Due to the differences in classes sizes, the combined trend (*i.e.*, ignoring classes) showed a positive change (*i.e.*, a typical FE) in scores of approximately 0.17 IQ points/year. Based on the GMM analyses, however, reporting the typical FE from the combined classes would be misleading, as there are actually two patterns of change over time. Specifically, Class 1 showed a decrease of approximately 0.07 IQ points/year, while Class 2 showed an increase of approximately 0.28 IQ points/year. Given the growth pattern of each class, the unconditional distribution of NIT scores at the third time point was expected to show negative skewness, which was supported and shown in Figure 4.

Table 7. Summary of model fit for the models with two–five-classes from the National Intelligence Test data.

	Fit Index							
Classes	**AIC**	**AICc**	**CAIC**	**BIC**	**SSBIC**	**DIC**	**ICL-BIC**	**LMR p**
2	11,043	11,019	11,097 *	11,086 *	11,051	11,091 *	11,301 *	0.09
3	11,029	10,997	11,102	11,087	11,040	11,094	11,361	0.60
4	11,023 *	10,983	11,116	11,097	11,036 *	11,105	11,439	0.34
5	11,024	10,976 *	11,136	11,113	11,040	11,124	11,544	0.49

Note. * Best fitting model. AIC, Akaike information criterion; AICc, corrected Akaike information criterion; CAIC, consistent Akaike information criterion; BIC, Bayesian information criterion; SSBIC, sample size adjusted-Bayesian information criterion; DIC, Draper's information criterion; ICL-BIC, integrated classification likelihood with Bayesian-type approximation; LMR p, *p*-value from the Lo–Mendell–Rubin likelihood ratio test.

Figure 3. Class-specific and combined growth patterns from matched respondents collected from three waves of Estonian National Intelligence Test scores.

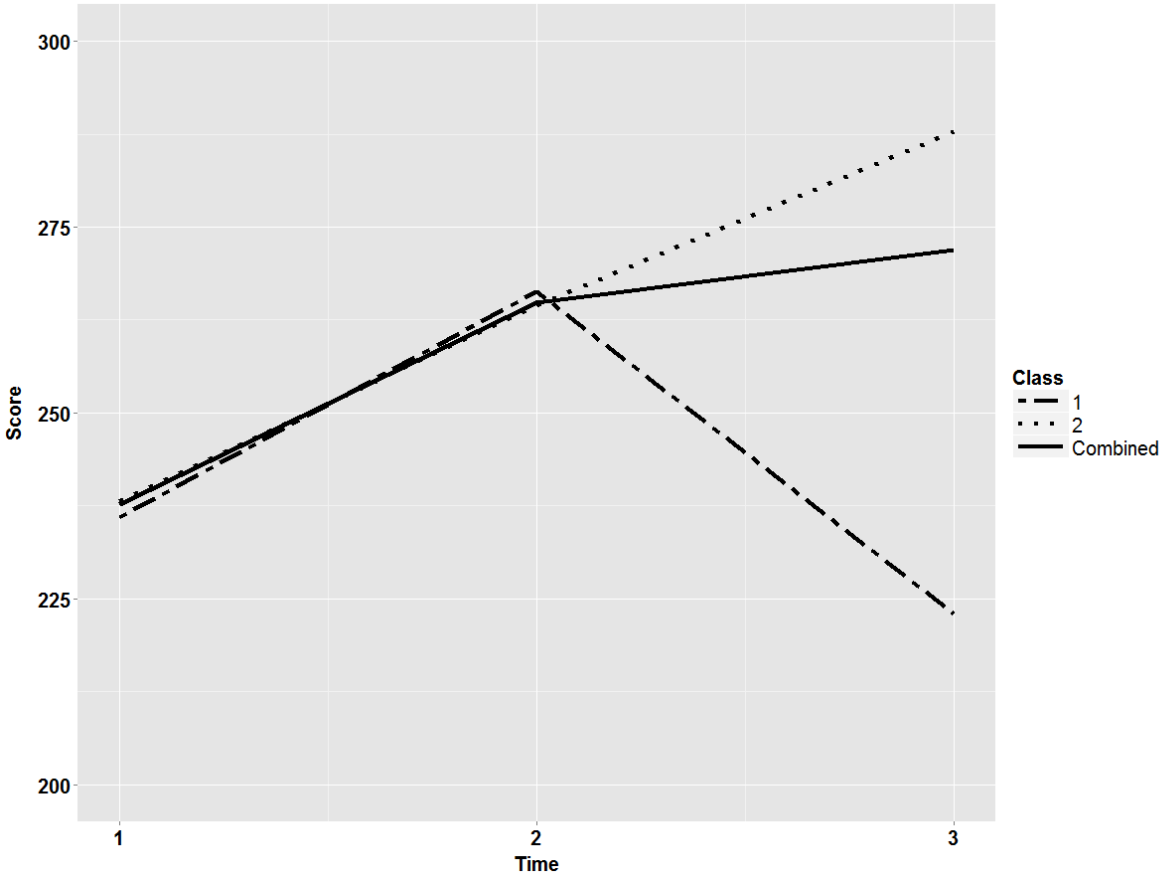
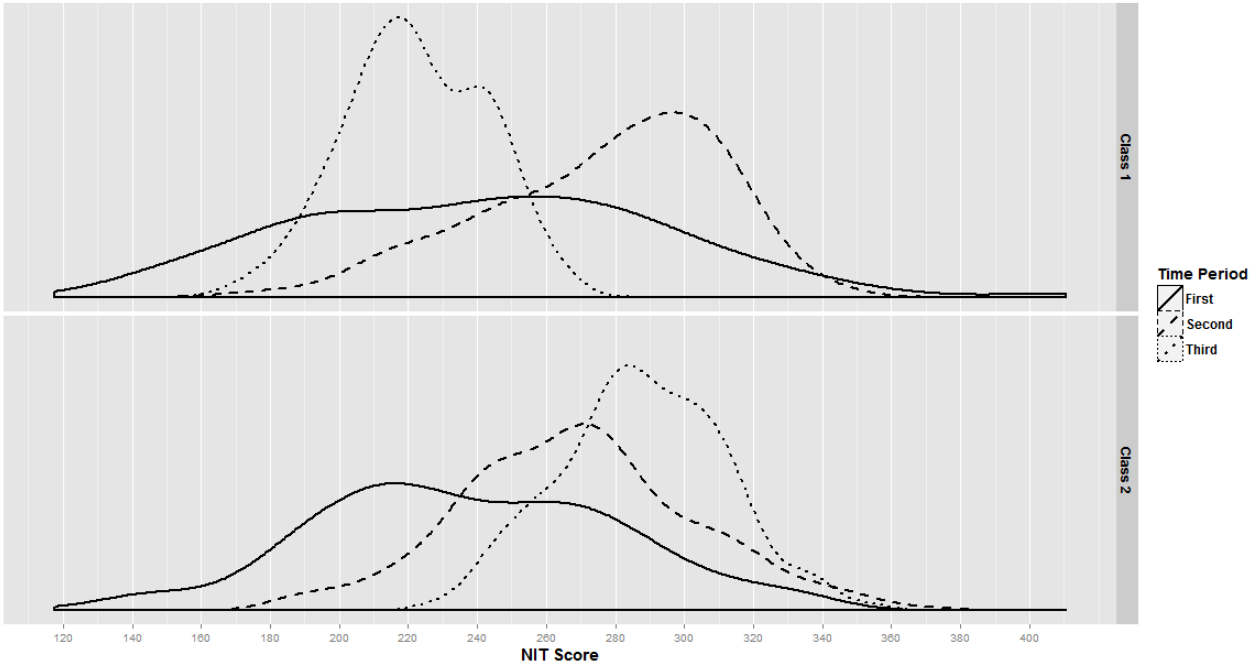


Figure 4. Distribution of National Intelligence Test scores for the three matched Estonian cohorts (1934, 1998 and 2006). The plotted score is the sum across the ten sub-tests.



4. Discussion

This study investigated the utility of commonly used indices of model fit in growth mixture modeling for examining the Flynn effect. Based on typical values from previous FE studies, we conducted a Monte Carlo (MC) study, simulating data from a two-class model that varied by four design factors: class prevalence rates, reliability of the measured variables, sample size and growth pattern shape (see Table 1). We then fit a series of two- to five-class models to the simulated data and examined eight fit indices (*i.e.*, AIC, AICc, CAIC, BIC, SSBIC, DIC, ICL-BIC and LMR) and the mean entropy for the two-class models to determine the relative accuracy in identifying a correct solution.

The results of the study suggest that when multiple classes underlie a given set of data, growth mixture modeling can be effectively used to uncover the number of unique growth patterns. That said, not all fit indices were equally able to detect the correct number of underlying classes. The CAIC and BIC were the most accurate at identifying the correct class solution. Furthermore, their performance was not negatively impacted by any of the factors included in the MC design. The DIC performed very well, too, but based on the current study's results, we cannot recommend its use over BIC, which, mathematically, is very similar to DIC and more commonly reported by latent variable modeling software. The ICL-BIC performed well, too, which was expected, given the performance of BIC and estimated entropy values. The performance of ICL-BIC will typically only deviate from BIC for models with lower estimates of entropy, because ICL-BIC is equal to BIC when relative entropy equals one. The SSBIC showed promise for identifying the correct number of classes when the sample size is sufficiently large (*i.e.*, when $n = 800$), which confirms previous research on this index [23]. Consistent with previous studies, the AIC did not perform well for model selection [29,48–52]. Given that relative entropy reflects the degree to which classes are distinguishable, we expected to observe the higher entropy estimates for the crossing growth pattern cells. Relative entropy estimates were considerably lower when the growth patterns were more similar (*i.e.*, more difficult to distinguish).

The LMR was the most sensitive fit index to the design factors of this study. Under the non-crossing growth pattern conditions, the LMR was very sensitive to the reliability of the growth indicators. The performance of LMR was also strongly impacted by the growth patterns underlying a dataset. That is, when the growth patterns exhibited a non-crossing pattern, the LMR correctly identified the number of classes in only about one-third of the datasets. When the growth patterns exhibited a crossing pattern, however, the LMR correctly identified the two-class solution in over 75% of the datasets.

4.1. Design Factors

The results of the MC study demonstrated the role that the underlying growth patterns play in the ability of certain fit indices to identify the correct number of classes. The crossing pattern was clearly less difficult to identify, because the classes' growth patterns are opposite of each other, whereas the non-crossing growth patterns only differed in the magnitude of the slopes (*i.e.*, both classes experienced a positive change over time, but to a different extent). In applied research, the expected class-specific growth patterns will be informed by the researcher's guiding theory and/or previous research. Therefore, these expectations should clearly be considered before using a GMM.

Contrary to our expectations, our study revealed that the reliability of the growth indicator variables did not generally have an adverse effect on the performance. These findings should be encouraging for applied researchers whose instruments' scores may not have a reliability of 0.95. We included an extreme (although not uncommon [68]) reliability condition to assess the degree to which fit was negatively impacted. For many indices, such as BIC, performance was very strong across both reliability conditions.

There was mixed evidence in this study on the effect of sample size. Sample size typically plays a very important role in mixture modeling, because the number of estimated parameters can increase drastically with just minor changes in the number of classes or model complexity. The models examined by this study were relatively simple models, which may have minimized the impact of the sample size in certain cases. Researchers estimating more complex models will likely need larger samples for convergence purposes.

4.2. Consequences of Incorrect Model Selection

As noted previously, model selection should incorporate substantive theory in addition to statistical criteria. This study focused primarily on statistical criteria that can be used to aid in model selection for FE studies, but in doing so has largely ignored the consequences of under- or over-estimation of the number of latent classes. The potential negative consequence(s) of interpreting an incorrect model depends on the context and the research question [69]. One of the benefits of identifying the number and characteristics of groups underlying a set of data is data simplification. When the number of classes interpreted is incorrect, then this benefit is jeopardized, because the group characteristics in the selected solution are not likely to approximate those in the mixture of populations. For example, if, when examining the Flynn effect, researchers incorrectly concluded that there were three classes when there was only one or if they treated an entire sample as being taken from a single population without giving consideration to the possibility of multiple underlying populations, then the decisions based on the models will be misleading. Furthermore, the larger the misfit between the true underlying model and the selected model, the more misleading the model-based conclusions. Unconditional distributions may appear questionable for various reasons, such as the existence of latent classes or simply poor instrument scaling [70]. Unfortunately, applied researchers cannot be certain the interpreted model is the true model, which is why considering multiple pieces of evidence is crucial to model selection and interpretation [71]. With the appropriate guiding theory and correct use of statistical fit indices, researchers may greatly improve the quality of their model selection, interpretations and conclusions.

4.3. Additional Considerations

Another consideration regarding the units to which the FE applies is also worth noting. The focus of the current study is on the potential utility and an the application of GMM for examining the FE. Our study examined the FE using data with two levels (*i.e.*, time nested within matched individuals), but GMM can also be used for analyzing data with more than two levels, such as when individuals are nested within a higher-level grouping variable (e.g., race/ethnicity, sex, socio-economic status). Likewise, combining GMM with genetically-sensitive research designs could help determine whether

the FE is the same between- and within-families [4]. Such extensions could prove useful in defining the different classes discovered by the GMM and, ultimately, in understanding the causes of the FE.

4.4. Comparison with Previous Flynn Effect Research Using the Estonian NIT

Scholars have previously used the Estonian NIT data to study the FE. Although they assumed only one growth pattern, they did find heterogeneity in the growth. Using different respondent subsets, Must, Must and Raudik [66] found aggregate score changes ranging from 0.04 (13âĂŞ14 year-olds) to 0.22 (12–13 year-olds) IQ points/year, while Must, te Nijenhuis, Must and van Vianen [65] found aggregate score changes ranging from 0.08 (14–15 year-olds) to 0.16 (13–14 year-olds) IQ points/year. Their reported score changes do not map directly onto the changes we found, but their results clearly support the idea of multiple patterns of change underlying the FE.

4.5. Recommendations

Based on the results of the current study, we make the following recommendations. First, FE scholarship should seek to gather data across multiple time periods. While there are some exceptions (e.g., [14,72,73]), most FE work only collects data from two time periods. The minimum number of time periods needed for a latent curve model is three, but four or more is better [17]; the same applies for GMMs. Second, given the heterogeneity of FE findings [4], more work should be done to determine the cause of this heterogeneity. For example, is it due to differences in the intelligence instruments, differences in the analyses, differences between countries or is it due to actual differences in the magnitude and direction of this secular trend? Our analysis of the Estonian NIT data, in conjunction with other analysis of the same data [65,66], suggest that age may be a factor in the FE. Nonetheless, more FE scholarship utilizing multiple time periods and appropriate data analysis models is needed. To that end, GMM can be a useful aid in better understanding the reasons for the heterogeneity. Third, if GMMs are used in subsequent FE studies, the results from the current study indicate that the CAIC and BIC fit indices will be useful model selection criteria across a variety of conditions.

Author Contributions

AB: Conceptualization; AB: Data Collection; GM: Programming & Simulation; GM: Statistical Analysis; AB, GM: Writing & Editing.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Flynn, J.R. *What is Intelligence? Beyond the Flynn Effect*; Cambridge University: New York, NY, USA, 2007.
2. Flynn, J.R. *Are We Getting Smarter? Rising IQ in the Twenty-first Century*; Cambridge University Press: New York, NY, USA, 2012.

3. Lynn, R. Who discovered the Flynn Effect? A review of early studies of the secular increase of intelligence. *Intelligence* **2013**, *41*, 765–769.
4. Williams, R.L. Overview of the Flynn effect. *Intelligence* **2013**, *41*, 753–764.
5. Rodgers, J.L. A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence* **1998**, *26*, 337–356.
6. Beaujean, A.A.; Osterlind, S.J. Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 Children and Young Adults data. *Intelligence* **2008**, *36*, 455–463.
7. Wicherts, J.M.; Dolan, C.V.; Hessen, D.J.; Oosterveld, P.; van Baal, G.C.M.; Boomsma, D.I.; Span, M.M. Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence* **2004**, *32*, 509–537.
8. Beaujean, A.A.; Sheng, Y. Examining the Flynn effect in the general social survey vocabulary test using item response theory. *Personal. Individ. Differ.* **2010**, *48*, 294–298.
9. Shiu, W.; Beaujean, A.A.; Must, O.; te Nijenhuis, J.; Must, A. An item-level examination of the Flynn effect in Estonia. *Intelligence* **2013**, *41*, 770–779.
10. Pietschnig, J.; Tran, U.S.; Voracek, M. Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypothesis perhaps. *Intelligence* **2013**, *41*, 791–801.
11. Cattell, R.B. The fate of national intelligence: Test of a thirteen-year prediction. *Eugen. Rev.* **1950**, *42*, 136–148.
12. Shiu, W.; Beaujean, A.A. *Evidence of the Flynn Effect in Children: A Meta-analysis*, Poster Presented at the Annual Meeting of the Association for Psychological Science, Boston, MA, USA, 2010.
13. Shiu, W.; Beaujean, A.A. *The Flynn Effect in Adults: A Meta-analysis*, Poster Presented at the Annual Meeting of the International Society for Intelligence Research, Arlington, VA, USA, 2010.
14. Kanaya, T.; Ceci, S.J.; Scullin, M.H. Age differences within secular IQ trends: An individual growth modeling approach. *Intelligence* **2005**, *33*, 613–621.
15. Dolan, C.V.; van der Maas, H.L.J. Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika* **1998**, *63*, 227–253.
16. Yung, Y.F. Finite mixtures in confirmatory factor-analysis models. *Psychometrika* **1997**, *62*, 297–330.
17. Bollen, K.A.; Curran, P.J. *Latent Curve Models: A Structural Equation Perspective*; Wiley: Hoboken, NJ, USA, 2006.
18. Ram, N.; Grimm, K.J. Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *Int. J. Behav. Dev.* **2009**, *33*, 565–576.
19. Rodgers, J.L. The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *Am. Psychol.* **2010**, *65*, 1–12.
20. Vermunt, J.K. *The Sage Encyclopedia of Social Sciences Research Methods*; Sage Publications: Thousand Oaks, CA, USA, 2004; chapter Latent profile model, pp. 554–555.
21. Rodgers, J.L.; Rowe, D.C. Theory development should begin (but not end) with good empirical fits: A comment on Roberts and Pashler (2000). *Psychol. Rev.* **2002**, *109*, 599–603.

22. Bollen, K.A.; Long, J.S. (Eds.) *Testing Structural Equation Models*; Sage: Newbury Park, CA, USA, 1993.
23. Tofighi, D.; Enders, C.K. *Advances in Latent Variable Mixture Models*; Information Age Publishing, Inc.: Greenwich, CT, USA, 2008; chapter Identifying the correct number of classes in growth mixture models, pp. 317–341.
24. Beasley, W.H.; Rodgers, J.L. Bootstrapping and Monte Carlo methods. In *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*; Cooper, H., Camic, P.M., Long, D.L., Panter, A.T., Rindskopf, D., Sher, K.J., Eds.; American Psychological Association: Washington, DC, USA, 2012; pp. 407–425.
25. Fan, X. Designing simulation studies. In *APA Handbook of Research Methods in Psychology (Vol. 3.): Data Analysis and Research Publication*; Cooper, H., Ed.; American Psychological Association: Washington, DC, USA, 2012; pp. 427–444.
26. Haggerty, M.E.; Terman, L.M.; Thorndike, R.L.; Whipple, G.M.; Yerkes, R.M. *National Intelligence Tests: Manual of Directions*; World Book: New York, NY, USA, 1920.
27. Bandalos, D.L.; Leite, W. *Structural Equation Modeling: A Second Course*; Information Age Publishing: Charlotte, NC, USA, 2013; Chapter: Use of Monte Carlo studies in structural equation modeling research, pp. 625–666.
28. Boomsma, A. Reporting Monte Carlo studies in structural equation modeling. *Struct. Equ. Model. A Multidiscip. J.* **2013**, *20*, 518–540.
29. Nylund, K.L.; Asparouhov, T.; Muthén, B.O. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Struct. Equ. Model.* **2007**, *14*, 535–569.
30. Kim, S.Y. Determining the Number of Latent Classes in Single-and Multiphase Growth Mixture Models. *Struct. Equ. Model.* **2014**, *21*, 263–279.
31. Singer, J.D.; Willett, J.B. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*; Oxford University: New York, NY, USA, 2003.
32. Lipsey, M.W.; Hurley, S.M. Design sensitivity: Statistical power for applied experimental research. In *The SAGE Handbook of Applied Social Research Methods*; Bickman, L., Rog, D.J., Eds.; Sage: Thousand Oaks, CA, USA, 2009; pp. 44–76.
33. McLachlan, G.J.; Peel, D. *Finite Mixture Models*; John Wiley & Sons, Inc.: New York, NY, USA, 2000.
34. Collins, L.M.; Lanza, S.T. *Latent Class and Latent Transition Analysis*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2010.
35. Bollen, K.A. *Structural Equation with Latent Variables*; John Wiley & Sons, Inc.: New York, NY, USA, 1989.
36. Lo, Y.; Mendell, N.R.; Rubin, D.B. Testing the number of components in a normal mixture. *Biometrika* **2001**, *88*, 767–778.
37. Vuong, Q.H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **1989**, *57*, 307–333.

38. Markon, K.E.; Krueger, R.F. An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behav. Genet.* **2004**, *34*, 593–610.
39. Muthén, B.O. LCA and Cluster Analysis. Message Posted to MPLUS Discussion List. Available online: http://www.statmodel.com/discussion/messages/13/155.html?1077296160 (accessed on 15 October 2013).
40. Vermunt, J.K.; Magidson, J. *Applied Latent Class Analysis*; Cambridge University Press: Cambridge, MA, USA, 2002; Chapter: Latent class cluster analysis.
41. Akaike, H. *On the Entropy Maximization Principle*; North-Holland: Amsterdam, The Netherlands, 1977; pp. 27–41.
42. Bozdogan, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **1987**, *52*, 345–370.
43. Hurvich, C.M.; Tsai, C.L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307.
44. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464.
45. Sclove, S.L. Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* **1987**, *52*, 333–343.
46. Draper, D. Assessment and Propagation of Model Uncertainty. *J. R. Stat. Soc. Ser. B (Methodological)* **1995**, *57*, 45–97.
47. Muthén, L.K.; Muthén, B.O. *Mplus: User's Guide*, 6th ed.; Muthén & Muthén: Los Angeles, CA, USA, 2010; pp. 197–200.
48. Celeux, G.; Soromenho, G. An entropy criterion for assessing the number of clusters in a mixture model. *J. Classif.* **1996**, *13*, 195–212.
49. Koehler, A.B.; Murphree, E.S. A comparison of the Akaike and Schwarz criteria for selecting model order. *Appl. Stat.* **1988**, *41*, 187–195.
50. Morgan, G.B. Mixed mode finite mixture modeling: An examination of fit index performance for classification. *Struct. Equ. Model.*, in press.
51. Soromenho, G. Comparing approaches for testing the number of components in a finite mixture model. *Comput. Stat.* **1994**, *9*, 65–82.
52. Yang, C.C. Evaluating latent class analysis models in qualitative phenotype identification. *Comput. Stat. Data Anal.* **2006**, *50*, 1090–1104.
53. Celeux, G.; Soromenho, G. An entropy criterion for assessing the number of clusters in a mixture model. *J. Classif.* **1996**, *13*, 195–212.
54. Henson, J.M.; Reise, S.P.; Kim, K.H. Detecting mixtures from structural model differences using Latent variable mixture modeling: A comparison of relative model fit statistics. *Struct. Equ. Model.* **2007**, *14*, 202–226.
55. Ramaswamy, V.; DeSarbo, W.S.; Reibstein, D.J.; Robinson, W.T. An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Mark. Sci.* **1993**, *12*, 103–124.
56. Muthén, B.O. *Mplus Technical Appendices*, Version 3rd ed.; Muthén & Muthén: Los Angeles, CA, USA, 2004; pp. 197–200.
57. Pastor, D.A.; Barron, K.E.; Miller, B.J.; Davis, S.L. A latent profile analysis of college students' achievement goal orientation. *Contemp. Educ. Psychol.* **2007**, *32*, 8–47.

58. Biernacki, C.; Celeux, G.; Govaert, G. Assessing a mixture model for clustering with the integrated classification likelihood. In *Technical Report*; Institut National de Recherche en Informatique et en Automatique: Rhone-Alpes, France, 1998.
59. McLachlan, G.J.; Ng, S.K. A comparison of some information criteria for the number of components in a mixture model. In *Technical Report*; Department of Mathematics, University of Queensland: Brisbane, Australia, 2000.
60. Bento, C.; A., C.; Dias, G. (Eds.) *Retails Clients Latent Segments*, Proceedings of the 12th Portuguese Conference on Artificial Intelligence, Covilha, Portugal, 5–8 December 2005; Springer-Verlag: Heidelberg, Germany.
61. Muthén, L.K.; Muthén, B.O. *Mplus: User's Guide*, 7th ed.; Muthén & Muthén: Los Angeles, CA, USA, 2012.
62. Hallquist, M.; Wiley, J. *MplusAutomation: Automating Mplus Model Estimation and Interpretation*, 2013, R Package Version 0.6-1.
63. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
64. Yerkes, R.M. *Memoirs of the National Academy of Sciences: Psychological Examining in the United States Army* ; Government Printing Office: Washington, DC, USA, 1921; Volume 15.
65. Must, O.; te Nijenhuis, J.; Must, A.; van Vianen, A.E.M. Comparability of IQ scores over time. *Intelligence* **2009**, *37*, 25–33.
66. Must, O.; Must, A.; Raudik, V. The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect. *Intelligence* **2003**, *31*, 461–471.
67. Schaie, K.W., Quasi-experimental research designs in the psychology of aging. In *Handbook of the Psychology of Aging*; Birren, J.E., Schaie, K.W., Eds.; Van Nostrand Reinhold: New York, NY, USA, 1977; pp. 39–58.
68. Peterson, R.A. A Meta-analysis of Cronbach's coefficient alpha. *J. Consum. Res.* **1994**, *21*, 381–391.
69. Millsap, R.E. Structural equation modeling made difficult. *Personal. Individ. Differ.* **2007**, *42*, 875–881.
70. Bauer, D.J.; Curran, P.J. Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychol. Methods* **2003**, *8*, 338–363.
71. McDonald, R.P. Structural models and the art of approximation. *Perspect. Psychol. Sci.* **2010**, *5*, 675–686.
72. Ang, S.C.; Rodgers, J.L.; Wänström, L. The Flynn effect within subgroups in the U.S.: Gender, race, income, education, and urbanization differences in the NLSY-Children data. *Intelligence* **2010**, *38*, 367–384.
73. Rodgers, J.L.; Wanstrom, L. Identification of a Flynn Effect in the NLSY: Moving from the center to the boundaries. *Intelligence* **2007**, *35*, 187–196.