DESIGNING A QUALITY MEASURE FOR BIRTH DELIVERY

by

Anna Streichhardt, B.A.

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Data Analytics and Information Systems
August 2021

Committee Members:

Emily Zhu, Chair

Lawrence Fulton

Rasim Musal

Tahir Ekin

Francis Méndez

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

Throughout my writing of this thesis, I have received a great deal of support and assistance. First of all, I would like to thank my supervisor, Dr. Emily Zhu, for her invaluable advice, continuous support, and patience from the beginning of my studies, throughout the thesis writing and completion. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I am deeply grateful for the P.E.O. sisterhood, who not only endorsed me as a role model for international female leadership, but also supported and sponsored my studies here at Texas State University through the international peace scholarship.

I would especially like to thank Dr. Lawrence Fulton, for providing me access to the Health Administration Lab and the department's lab machine which massively helped my computations. Thank you for your patient support in the process of getting access to the lab. Furthermore, I would like to thank Dr. Rasim Musal for taking his time to explain MCMC theory to me and all of his very helpful comments and suggestions. Additionally, I would like to thank Dr. Francis Méndez and Dr. Tahir Ekin for being part of my committee and their insightful comments, feedback, and suggestions.

I would like to thank Dr. Jaymeen Shah, for always having an open ear and his continuing support from the beginning of my admission at Texas State University, and always finding answers to all of my questions, no matter what.

A special thanks goes also to Dr. Alexander McLeod and his mentorship. You not only supported me as an exchange student from the beginning of my journey at Texas

State University, but also made pursuing my master's here at Texas State possible. My appreciation also goes out to Dr. Li Feng and our research team for always supporting and looking out for me.

I want to give special thanks to my partner Roberto D. Guerra and his family, who supported me in my thesis and helped me in every way to succeed my studies. Additionally, I would like to offer my special thanks to Dr. Jaime Rodriguez and Dr. Miriam Guerra for their valuable help in proofreading my thesis.

Finally, a special thanks goes also to my family and friends, who did all they could to help me in my professional and personal development even if they were across the world. Thank you for your unwavering support and belief in me.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Abbreviation	Description
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CFI	Comparative Fit Index
CI	Confidence interval
CPR	Cardiopulmonary resuscitation
GHDM	Generalized heterogeneous data model
HMC	Hamiltonian Monte Carlo
IOM	Institute of Medicine
IRT	item response theory
LOOIC	Leave-one-out cross validation
MCMC	Monte Carlo Markov Chain
NICU	Neonatal Intensive Care Unit
mPCA	mixed Principal Component Analysis
PCA	Principal Component Analysis

RMSE	Root Mean Square Error
RMSEA	Root Mean Square Error of Approximation
SD	Standard deviation
SRMR	Standardized Root Mean Square Residual
SS	Sum of squared factor loadings
TLI	Tucker-Lewis Index
WHO	World Health Organization
WSS	within groups sum of squares

1: INTRODUCTION

1.1. Motivation and Background

The World Health Organization (WHO) defines health as "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity" (World Health Organization, 2004b). Thus, health is a complex interaction of multiple internal and external factors. The Institute of Medicine (IOM) reports an urgent need to address underlying factors that jeopardize health and that are not limited to disease outcome. To detect and measure health outcome indicators, researchers provide a clear understanding of the current situation and needed advancements in government and healthcare industry to improve health (Koh, 2011). An interest for health care indices has increased globally, but according to (Murray & Frenk, 2008), the goal of improving health services is impossible without proper measurement, systematic evaluation, and data analysis.

The WHO defines quality of care as "the extent to which health care services provided to individuals and patient populations improve desired health outcomes. In order to achieve this, health care must be safe, effective, timely, efficient, equitable, and people-centered" (World Health Organization, 2004b); however, the WHO does not provide a measurement or quantification to assess quality of care.

In 2016, the WHO published new standards to improve maternal and neonatal care quality in healthcare facilities. The core concept of these standards is focused on people and their experiences with health care. According to the WHO, the central principles of maternal care include competent and motivated health-care professionals,

effective communication based on women's needs, and community engagement.

When severe events happen (i.e., maternal complications, newborn abnormalities, etc.) during birth delivery, the healthcare system is burdened with providing extra care for long-term complications impacting mother's and/or child's quality of life. This burden is immediate, increases healthcare costs, and leads to potential bottlenecks in emergency treatments. The lack of quantifying maternal care quality makes it difficult for politicians to create new policies and to address current issues. Thus, it underlines the need for quality quantification that aims to decrease morbidity rates, raise the quality of maternity care, and improve hospital planning (Collier & Molina, 2019; Health Affairs Blog, 2019).

Maternal quality of care is crucial for the mother's and newborn's well-being. In 2018, for every 100,000 live births in the United States, 17 maternal deaths occurred. This rate is about double when compared to Australia, Canada, France, Germany, the Netherlands, New Zealand, Norway, Sweden, Switzerland, and the United Kingdom. According to (Roosa Tikkanen, 2020), 17% of maternal deaths occurred on the day of birth delivery. These mortality rates have led to an increased focus on maternal care in the United States with a need for standardization and improvement of maternal quality and safety using data-driven recommendations for system improvement (Collier & Molina, 2019).

A current health measure for the newborn's health is the Apgar score. This score is assessed immediately after the birth delivery. The Apgar score is currently used by insurance companies, physicians, hospital, and parents to provide an immediate clinical assessment of the newborn's condition (American College of Obstetricians and

Gynecologists, 2015; Siddiqui et al., 2017). Although the Apgar score is a standard procedure in the United States, it is considered a poor predictive value for the long-term outcome of the newborn because it only assesses the clinical status of a newborn and does not include the mother's status. Furthermore, a resuscitated newborn's score differs greatly from that of an autonomously breathing newborn (American College of Obstetricians and Gynecologists, 2015). Thus, this study aims to incorporate the mother's and newborn's clinical status immediately after birth delivery into one single score.

It is crucial for proper hospital management, emergency planning, and policy-making to have sufficient birth quality metrics to make proper decisions affecting long-term care and quality improvement. Maternal care processes vary by 50-100% in a single hospital. Neither patients nor hospital could explain those inter-hospital variations (Kozhimannil, 2014). By allowing a direct comparison of individual birth delivery cases, possible reasons for this inter-hospital variations and differences in care processes can be identified and addressed, the variability of care processes can be captured and effectively improved to provide a standardized health measure for birth delivery quality. Furthermore, the quality measure can answer quality questions concerning governing processes. For example, do certain physicians or hospitals get the more "serious" and "complicated" cases? Are there trends and patterns in patient characteristics across regions and areas? This study posits a quantitative quality measure that could be used to answer these questions and furthermore allow for adjustments in maternal care expectations and planning. The proposed quality measure could be further used for bench-marking and hospital comparisons (Kozhimannil, 2014). Hospital planning policies usually have a specific goal, for example, reducing the maternal death by 5% in

the next two years. However, such concrete policy-making requires measurable metrics that so far have not been established in the maternal care quality. The lack of birth delivery quality measures causes gaps in the improvement of maternal care. Policy-making and long-term planning fail to be efficient if the underlying quality cannot be captured in health metrics (Moller et al., 2018).

1.2. Data Description

The dataset used is the 2018 Natality public use data (further referred to as Vital Birth Records 2018 or dataset) obtained from the National Center for Health Statistics (NCHS) provided through the Center of Disease Control (CDC) and includes more than 99% of all live births of birth deliveries in the United States in 2018 from citizens and non-citizens.

The dataset includes continuous variables about the mother's age at birth, the birth weight in grams, and the gestation time of the pregnancy in weeks (ranking from 17 – 47). The Apgar scores (measured for the newborn after 5 and 10 minutes) are also available in the dataset. According to ICD-10 low birth weight is defined as less than 2,500 grams. The average birth weight for babies born between the 37^{th} and 40^{th} gestation week (referred as full-term) is 3,200 grams (University of Rochester Medical Center Rochester, 2020). Apgar scores are standardized measures used to quantify the newborn's fitness after it is born. The Apgar score 5 is measured between 1 and 5 minutes after the birth delivery. It is composed of 5 components: 1) the newborn's ability to breathe; 2) heart rate; 3) muscle tone; 4) grimace response or reflex irritability; and 5) skin color. On every section, the newborn can score a maximum of 2 points (American College of Obstetricians and Gynecologists, 2015). If the 5-minutes Apgar score (further

referred as Apgar 5) is 5 or less, the Apgar score is measured again after 10 minutes (further referred as Apgar 10) (National Center for Health Statistics, 2019).

The dataset includes the following information about *Maternal Morbidity* (if occurred or not), including maternal transfusion, perineal laceration, ruptured uterus, unplanned hysterectomy, and admission to intensive care. It includes 6 variables for Abnormal conditions of the newborns (event occurred or not) describing assisted ventilation immediately after delivery, assisted ventilation required for more than 6 hours, admission to NICU, newborn given surfactant replacement therapy, antibiotics received by the newborn for suspected neonatal sepsis, seizure or serious neurological dysfunction. Additionally, the dataset includes 12 variables about genetic anomalies categorized as *Congenital abnormalities of the newborns* (diagnosed or not diagnosed) which include anencephaly, meningomyelocele / spina bifida, cyanotic congenital heart disease, congenital diaphragmatic hernia, omphalocele, gastroschisis, limb reduction defect, cleft lip with or without cleft palate, cleft palate alone, Down syndrome, suspected chromosomal disorder, and hypospadias. All of the 24 listed variables are recoded as yes or no (binary) and in general occur rarely. More than 90% of all birth deliveries in 2018 do not exhibit any of these severe conditions. However, they severely affect the overall quality of life of the newborn or, if an event of the maternal morbidity occurred, severely impact the mother's health. Descriptive statistics of the dataset and a short description of the relevant complications are presented in Chapter 3.

1.3. Contribution

The purpose of this proposed thesis is to design a quality measure for individual birth delivery outcomes that considers the health status of mother and child. Quality of

care is the key component of healthcare and the equity and integrity for woman and child. In order to achieve and maintain a good quality of care, quantifying the quality of care is crucial. Not only gives it an overview over the current status of the healthcare system, it also allows to create performance benchmarks, and peer comparisons in order to improve quality (McDowell et al., 2004).

So far, health metrics for maternal care focus on mortality, not on morbidity statistics (Health Affairs Blog, 2019). This thesis aims to close the gap between the understandings of maternal care, more specifically birth delivery, and the quantification of its quality. Hereby, birth delivery quality includes mother as well as newborn(s) as one entity. Mother's and child's health are likely strongly correlated since the baby is part of the mother's body until birth delivery, and therefore dependent of the mother's overall health status. By creating a health index measuring the health status of newborn and mother, the study is aiming to provide a better understanding about the quality of birth delivery in the United States by making a complex definition as quality tangible to even non-healthcare professionals. Furthermore, this study sees mother and child as one entity instead of focusing on one or the other, which creates a stronger overall quality picture. To the author's knowledge, there has been no latent variable model using mixed response variables using a latent continuous factor as health indicator in the birth delivery area. To fill this gap, this study proposes a birth delivery measure which quantifies individual birth deliveries by taking mother's and child's condition into account. The thesis is structured as follows: the literature review about the Apgar score and different methods to estimate a latent variable with mixed data types is presented in Chapter II. Chapter II also discuss common software and available packages, and provides background for performance

measures. Chapter III presents an overview of the dataset and provides more detail about our data cleaning and transformation process. Chapter VI, presents the model results, naming factor analysis, structural equation modeling, Hamiltonian Monte Carlo, and clustering. A comparison of the different model types and verification of the best overall model with birth delivery data from 2017 is also presented. Chapter V presents a discussion of the findings, limitations of the study, and recommendations for future research.

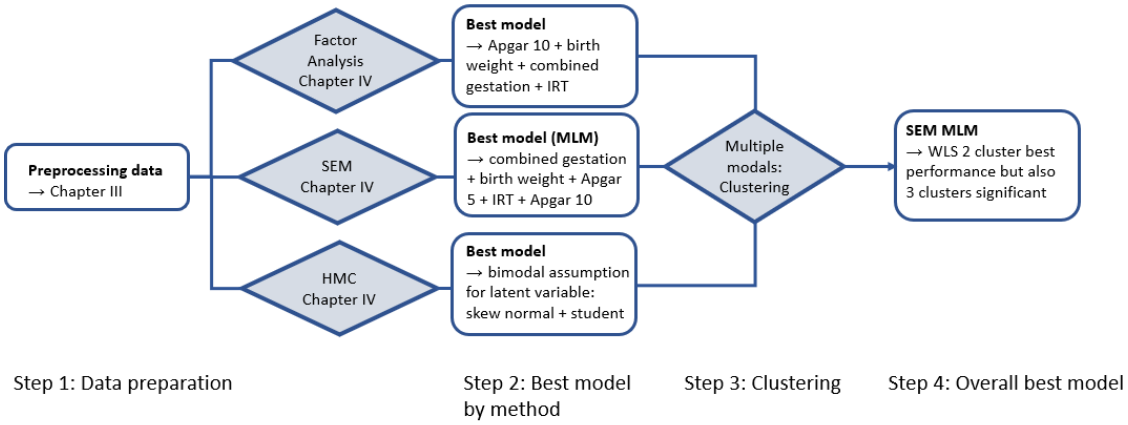

Figure 1: Summary of applied methods and choices

2: LITERATURE REVIEW

2.1. Apgar Score

An important health measure included in our analysis is the Apgar score. The score has been developed in 1952 by Dr. Virginia Apgar to answer the need of an immediate assessment of the newborn's clinical status after birth. The standardized Apgar score ranges from 0-10, where 7-10 is considered normal, 4-6 moderately abnormal, and 0-3 as abnormal. Newborns which receive an Apgar score of 0 have usually very little chance of survival. If they survive, usually not without neurological damage. However, Apgar scores cannot be used to predict an individual's adverse neurologic outcome. Low Apgar scores are correlated with low birth weight. Nonetheless, the Apgar score alone is no stand-alone predictor for a baby's mortality and morbidity. Table 1 represents the chart a physician has to fill to calculate the Apgar score (American College of Obstetricians and Gynecologists, 2015; Reiter & Walsh, 2017).

Table 1: Apgar score chart

Indicator		0 Points	1 Point	2 Points
A	**Activity (muscle tone)**	Absent	Flexed arms and legs	Active
P	**Pulse**	Absent	Below 100 bpm	Over 100 bpm
G	**Grimace (reflex irritability)**	Floppy	Minimal response to stimulation	Prompt response to stimulation
A	**Appearance (skin color)**	Blue; pale	Pink body, Blue extremities	Pink
R	**Respiration**	Absent	Slow and irregular	Vigorous cry

The Apgar score is found not only on a national level, some European countries or regions record them as well. Siddiqui et al. (2017) conducted a study to test its

reliability across twenty-three countries and regions including includes 2,183,472 live births. For their study, they aggregated and investigated Apgar 5 scores from the Euro-Peristat project in 2004 and 2010. They uncovered large variations across different countries and regions, majorly based on different assessment by nation. This difference may be caused by differences in wordings of national guidelines (Siddiqui et al., 2017).

2.2. Search Methods

As a first step, we searched on google scholar with the following search terms: latent variable on mixture data, dichotomous/binary and continuous observed variables and continuous latent variable, factor analysis on mixture data with skewness, and latent factor analysis on mixed data dichotomous and continuous variables. Other search terms were dichotomous observed variables, continuous observed variables, continuous latent variable, latent factor analysis, latent variable analysis, mixed data. The first 10 pages were searched, if relevant papers were found after page 9, we searched until page 15.

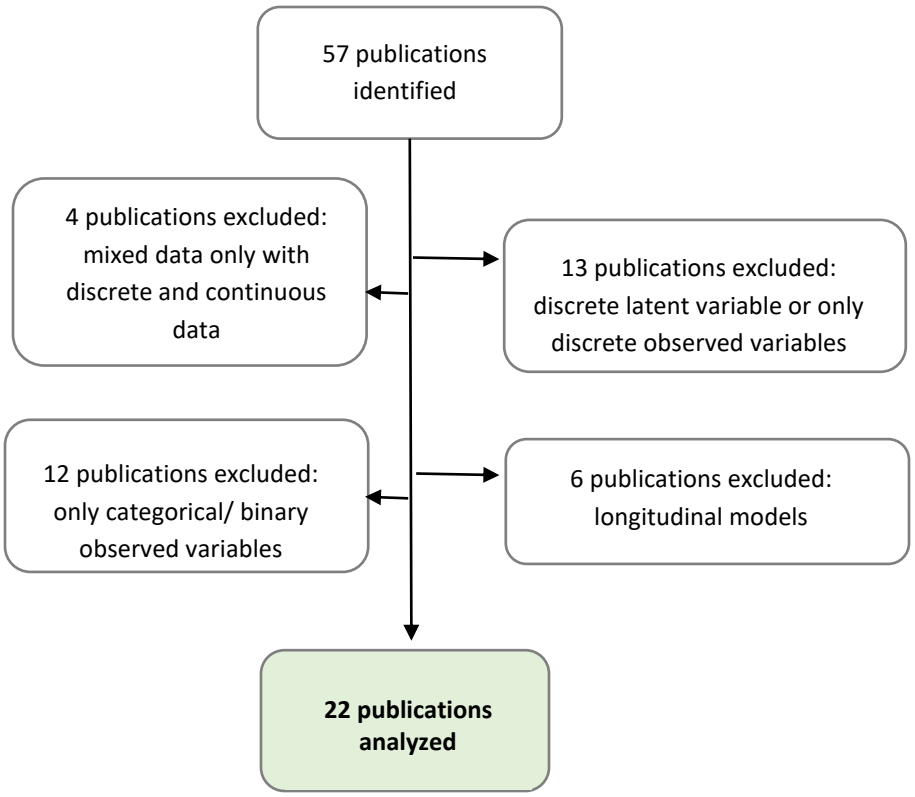

Figure 2: PRISMA flow diagram of literature search

In total, 57 publications were identified in the first step as illustrated in Figure 2. The publications were searched for the following keywords: continuous, categorical/ ordinal, and latent. We excluded publications with models for longitudinal data and publications with models without mixed data types. Furthermore, we excluded publications where models did not include or were just restricted to categorical observed variables. After applying the exclusion criteria, 22 resulted as relevant. In a second step, cited by references of the found relevant literature in google scholar were analyzed. A maximum (if applicable) of 100 citations were explored. In a third step, publications published in a journal with an impact factor higher than 3.5 were analyzed for potential relevant literature. Overall, 55 more relevant publications were found in the last two steps resulting in a total of 77 relevant publications dealing with latent variable models with

mixed data.

2.3. Latent Variable Models

In the recent years, the interest for mixed latent models has immensely increased. This is owed to recent technology improvements leading to higher computational speed for more complex models. With the era of big data more complex data structures are available which created a need for more complex models (Bhat, 2015). Latent variable models using mixed data (binary, ordinal, continuous variables) are frequently used in socioeconomics, sociology, biology, and psychology but also economics, finance, transportation, and various other fields (de Leon & Chough, 2013). Their purpose is to address hidden variables which are not measured directly but indirectly through relevant variables in the dataset (Bauer & Curran, 2004; Ramezani et al., 2019; Varriale & Vermunt, 2012). I will refer to the variables in the dataset which are associated with the latent variable as response variables, observed variables or indicators.

2.3.1. Structural Equation Modelling (SEM)

One possible method for latent variable modeling with mixed data is structural equation modeling. As one of the early researchers of latent variable models with mixed categorical, dichotomous, and continuous indicators Muthén described a Structural Equation Model (SEM) approach where he combined mixed data using a probit regression model (Muthén, 1984). As one of the leading pioneers for latent variable models, Jöreskog described with his work first equation models in the field of psychology creating a base for further research (Jöreskog, 1977). The continuous latent variable formed by mixed responses underlies the normality assumption in most structural

equation models as shown in many examples (Bhat, 2015; Bhat et al., 2016; Muthén, 1984; Rabe-Hesketh et al., 2004; Ramezani et al., 2019; Shi & Lee, 2000; Song & Lee, 2006; Wang et al., 2020; Yuan, 2016; Yuan & Chan, 2016; Yuan et al., 2011; Zhou et al., 2014).

2.2.1.1 Skewed variables in SEM

One way to deal with nonnormal variables in SEM applications is to apply normal transformations in order to ensure the normality assumption. Wang et al. use the log transformation on the observed nonnormal variables (Wang et al., 2020), Ramezani et al. (2019) apply the Kaiser normalization on the indicators (Ramezani et al., 2019), while Bhat (2015) normalizes the covariance matrix (Bhat, 2015) before estimating the model parameters. For SEM with moderately skewed variables, Yuan et al. suggests the usage of the maximum likelihood with Satorra–Bentler scaled statistics (Yuan et al., 2011). For ordinal and continuous data with unknown population distribution in SEM is unknown and the observed variables show skewness, Yuan and Chan (2016) recommend to use the ridge Generalized Least Squares after replacing the covariance matrix with the polychoric, polyserial, and/or Pearson correlation matrix (Yuan & Chan, 2016). Muthén and Santorra (1995) warn about the convergence rate for asymptotic distributions using Generalized least square estimates for large models (Muthén & Satorra, 1995). Rabe-Hesketh et al. (2004) introduced a generalized linear latent and mixed modeling (GLAMM) framework to unify multilevel SEMs where they indicate the use of the nonparametric maximum likelihood distribution for nonnormal latent variables (Rabe-Hesketh et al., 2004).

2.2.1.2. Bayesian SEM

Another variation of SEM is the Bayesian SEM which has become increasingly popular in SEM calculations because of its effectiveness in complex SEMs and multilayered data structures. Unknown model parameters are treated as random variables and then evaluates posterior distributions (Adnan & Thanoon, 2015). Typically a Bayesian SEM approach requires normality assumptions. Adnan and Thanoon examine SEM using mixed models in a survey and concludes that parametric SEMs in many times fail to capture subtle data pattern across the predictors. Furthermore they describe the issue that SEMs are traditionally designed as parametric framework and unable to incorporate mixed data types. They recommend the use of generalized and semiparametric SEMs as described by Song et al. (Adnan & Thanoon, 2015; Song et al., 2013).

To handle nonnormality in SEMs, Lee and Song developed a Bayesian approach in the context of a general nonlinear structural equation model by including a normalizing constant (Lee & Song, 2004). They further developed Bayesian modeling approach for generalized semiparametric Structural Equation Models handling mixed data types. This approach uses Bayesian P-splines and an improved Markov chain Monte Carlo (MCMC) algorithm to estimate the underlying function. For model selection, they introduce the deviance information criterion which is defined as Bayesian statistic for model selection (Adnan & Thanoon, 2015; Song et al., 2013). Song et al. (2013) published their created R code on their website but unfortunately it is not available anymore. Yang and Dunson (2010) also extend the SEM to a Bayesian semiparametric SEM by implementing a centered Dirichlet process combined with a Markov chain Monte Carlo algorithm in their

approach which allows the latent variable to have an unknown distribution. They create their model using Fortran code but their code is not available to the public anymore (Yang & Dunson, 2010).

As a special SEM model, Fahrmeir and Raach propose a Bayesian semiparametric latent multiple indicator multiple cause (MIMIC) model using binary, ordinal and continuous responses. While having a data pool of 170,000 observations, they use a subsample of 6,804 observations for their model consisting of 2 latent variables with 10 indicators and refer to it as large sample size (Fahrmeir & Raach, 2007).

2.3.2. Mixed Factor Analysis

As a second method on how to combine mixed data types in a latent variable model Factor analysis was first introduced by Spearman (1904) with further development by Thurstone (1935,1947) who principally focused on continuous factors (Bauer & Curran, 2004). In the more recent years, the term factor analysis has also been used for models with binary or ordinal indicators. Mixed factor analysis describes a framework which combines the item-response theory model for dichotomous data and the normal factor model for continuous or ordinal data using a probit link function. Mixed factor analysis models are commonly used in models with mixed data types (Bauer & Curran, 2004; Lee & Shi, 2001; Tillmann et al., 2020; Wendt et al., 2019). One of the underlying assumptions of factor analysis models is the independency assumption of the observations (Varriale & Vermunt, 2012). In general, Gaussian or multivariate Gaussian distributions still are most commonly used in latent variable models to develop the continuous latent variable (Nussbaum & Giesen, 2020). The factor analysis model returns factor loadings for the estimation of the latent variable which can further be used in either

MCMC estimation or EM algorithm to estimate the probability distribution of the latent variable (Gruhl et al., 2013; Nussbaum & Giesen, 2020; Rosenthal & Voeten, 2007). MCMC and EM algorithm will be explained later in this chapter. Feng et al. (2017) introduce a generalized confirmatory factor analysis model for handling mixed data which assumes normality of the latent variable. The model utilizes the correlation matrix to combine the indicators. Distribution estimation is created by using a MCMC algorithm (Feng et al., 2017). Quinn introduces a Bayesian Factor Analysis for mixed ordinal and continuous responses, where the model generalizes both standard normal theory factor analysis models and item response theory models for ordinal data using a straightforward MCMC method. (Quinn, 2004) This method has been applied by McManus and Nieman using 12,366 observations (McManus & Nieman, 2019). Sánchez develops a Bayesian factor analysis model using 14 mixed ordinal and continuous indicators to create an index utilizing a MCMC algorithm for distribution calculation. He incorporates the internal correlation of the indicators and uses WinBugs for computation. Unfortunately, he does not indicate the number of observations so it is not clear if his model could run on a huge dataset (Sanchez, 2014). Meijer et al. create a health index and therefore introduce a LISCOMP model which integrates factor analysis and regression based on Muthen and Santorra's work for mixed data types. The model assumes normality which causes issues by including binary variables with low frequencies in distribution estimation methods. As a solution, they derived the full information joint loglikelihood function (Meijer et al., 2011). Nussbaum and Giesen developed traditional factor models further and introduced a pairwise sparse and low-rank model for mixed variables as a novel approach. The pairwise sparse interaction can be seen as the common idea in factor models measuring

correlation, the low-rank part can be seen as the effect of the continuous variables which are assumed to be conditional Gaussian. The general idea is to address the joint model parameters as a convex regularized likelihood optimization problem. An advantage of this approach is its consistency even for high dimensional models. The model is written in python and made as a python package publicly available on GitHub (Nussbaum & Giesen, 2020).

The normal maximum likelihood is not applicable with skewed data. Wall et al. (2012) use a mixture factor analysis model with skewed latent factors and multiple sample sizes whereas the largest sample size has 1,000 observations. They conduct a simulation study to compare normal maximum likelihood and mixture factor analysis and discover that if the sample size contains 500 observations or more, the mixture factor analysis is nearly unbiased by the skewness while the normal maximum likelihood stays biased (Wall et al., 2012).

2.2.2.1. Mixed Factor Analysis with skewed variables

To deal with skewed variables researchers developed multiple approaches. Wall et al. (2012) introduce a mixture factor analysis assuming a mixture of normal distributions (Wall et al., 2012). Murray et al. (2013) introduce a modified Bayesian mixed factor analysis by extending it to a semiparametric approach. They created an R package bfa which has been removed from the R repository (Murray et al., 2013). Gruhl et al. (2013) develop a semiparametric latent variable model with an extended rank likelihood estimation which they apply in a Bayesian factor analysis. The approach avoids the specification of a distribution for the indicators given the latent variables can be applied in SEM or other latent variable models. However, this approach did not test

formal model selection criteria (Gruhl et al., 2013). Feng et al. (2019) develop a joint modeling framework factor analysis model to create a latent financial literacy index. Their joint model operates with the underlying normality assumption but outperforms conventional models. It addresses potential multicollinearity issues, and reduces the dimensionality for the novel latent two-part regression model. They introduce their model using the open-source software Stan which is a programming language used for Bayesian analysis with an advanced MCMC algorithm, the so-called Hamiltonian Monte Carlo method (HMC), which provides fast computation. Stan is available through an R interface but the HMC method is in general more complicated than conventional methods and takes significantly more time to develop (Feng et al., 2019).

2.3.3. Moment Tensor Approach

A third way to estimate latent variables is proposed. Zhao et al. develop a generalized method of moments (GMM) approach for fast parameter estimation which extends current moment tensor approaches. Zhao et al. (2018) prove that their GMM approach utilizing moment tensor methods shows advantages compared to MCMC and EM methods. They proposed a python package called MELD which describes a generalized Dirichlet latent variable model utilizing a moment tensor approach for parameter estimation (Zhao et al., 2018). For big sample sizes this approach may become problematic. It utilizes the Newton-Raphson method which requires high computational capacity and therefore limits its use for big datasets (Civek & Kozat, 2017). However, no documentation or vignettes is available about proper use for binary variables, and no example with only continuous or mixed binary-continuous given which leads to errors when trying to apply our dataset.

2.3.4. Generalized Heterogeneous Data Model (GHDM)

A fourth method for latent variable models is described by Bhat et al. They introduce a so-called Generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. It models the covariance relationships among the mixed data types resulting in continuous latent factors. The model consists of two parts: a latent SEM and the latent variable measurement equation model (Bhat, 2015; Bhat et al., 2016). The GHDM does not have any available packages in R. We will skip this method and move forward with factor analysis and SEM.

2.3.5. General Mixed-Data Model

Samani and Ganjali (2010) introduce a joint model utilizing a general location model for ordinal and continuous responses. Their model assumes multivariate normal distribution of observed variables which presume a linear correlation with the latent variables (Samani & Ganjali, 2010). A general mixed-data model has also been introduced by de Leon and Carriègre (2007). They combine a general location model which is used for joint modeling for continuous and nominal variables with a conditional grouped continuous model which uses a latent variable which utilized the multivariate normal distribution (de Leon & Carriègre, 2007; Tabrizi et al., 2020). Paleti, Bhat & Pendyala (2013) extended their model so it could include mixed continuous, nominal, ordinal, and count variables assuming multivariate normal distribution (Paleti et al., 2013; Tabrizi et al., 2020). Amiri et al. proposed a mixed latent variable to jointly include nominal, ordinal, count and continuous data considering Poisson and normal distribution for the observed variables (Amiri et al., 2018). Taking the approached from Paleti, Bhat & Pendyala (2013) and Amiri Khazaei, & Ganjali (2018) into account, Tabrizi et al.

(2020) introduce a general latent model for mixed continuous and categorical indicators which extends de Leon and Carriere's model. Their model uses a joint density function in conjunction with conditional and marginal normal distribution (Tabrizi et al., 2020). The General mixed-data model does currently not have any available packages in R. We will skip this method and move forward with factor analysis and SEM.

2.3.6. Other Methods for Latent Variable Estimation

Gaussian copula models are constructed utilizing the normal or multivariate normal distribution and have been extended to fit mixed data types. They mainly rely on EM or MCMC methods which makes them computationally intensive (Zhao et al., 2018). Murray et al. (2013) developed a semi-parametric Bayesian Gaussian copula model using the extended rank likelihood (Murray et al., 2013). Another copula approach is carried out by Jiryaie, Withanage, Wu, & de Leon (2016) by introducing a class of mixed-variable distributions generated by Gaussian copulas (Jiryaie et al., 2016). Jafari, Tabrizi, & Samani (2015) propose a copula-based regression model to handle mixed ordinal and continuous indicators assuming underlying normality (Jafari et al., 2015). Using a Gaussian Copula-based regression model Rezai Ghahroodi et al. incorporate their mixed data model (Rezaei Ghahroodi et al., 2019). Cui (2019) proposes a Bayesian Gaussian copula model combined with robust maximum likelihood (Cui, 2019). More recently, Zhao proposed a semiparametric algorithm models mixed data using a Gaussian copula model which models multivariate distributions via transformations of a latent Gaussian vector. This approach is proposed as a missing value imputation which is of less interest for this study (Zhao & Udell, 2020).

However, there are several disadvantages of copula models. Most importantly,

copula models lack in robustness for highly correlated variables and therefore lead to over-dispersion. Furthermore, they are associated with less efficiency and slow computation time. Zhao et al. (2018) compare their moment tensor approach to the Bayesian copula factor model and notice much higher error rates for the copula models (Marchese, 2018; Zhao et al., 2018).

Another method which is of less interest for this study is the approach to create subgroups based on clustering methods using the binary variable(s) of the dataset. Then, the latent variable for each subgroup is estimated (Bauer & Curran, 2004; Depaoli & Clifton, 2015; Kelava & Brandt, 2014; Lin et al., 2016; Lubke & Luningham, 2017; McParland et al., 2017; Morlini, 2012; Rabe-Hesketh & Skrondal, 2006; Ranalli & Rocci, 2017; Tillmann et al., 2020; Varriale & Vermunt, 2012). Since our dataset includes 23 binary variables for birth delivery complications this would go beyond the scope of a feasible approach.

2.4. Density Ratio Model

A fifth method to estimate a hidden variable is described in Marchese's work. Marchese (2018) proposed an alternative to latent variable models: a density ratio model utilizing a semiparametric regression model for outcomes from an exponential family of distributions. In his work, he develops a general joint regression model approach which includes binary, count and continuous data with a maximum likelihood function for parameter estimation. To implement his approach, he uses R. However, when testing his model on a simulation, he states a limitation of his work to medium sample sizes (Marchese, 2018).

2.5. Parameter Estimation

2.5.1. Markov Chain Monte-Carlo (MCMC)

MCMC is a computer-driven sampling method which is used for distribution calculation when parameters are unspecified. An advantage of MCMC is that the distribution calculation is possible even when only the density calculation for random samples is available. The MCMC method draws random samples from a distribution which does not need to be a normal distribution. It is often used in Bayesian frameworks to estimate parameters. The MCMC algorithm works as follows: First, a so-called proposal for the newly drawn sample is calculated with including a small random deviation compared to the sample before. Then, this calculated proposal is either accepted or rejected. If accepted, the random sample becomes part of the MCMC sample chain which completes one iteration of the method. With enough samples the iteration stops. One drawback of this method is the so-called burn-in: if the initial sample was supposed to be very unlikely considering the target distribution, the method uses way more iterations than originally needed. This means starting points are crucial for the algorithm's burn-in and practical convergence (van Ravenzwaaij et al., 2018).

MCMC is available in R under the nimble package and the MCMC package (de Valpine et al., 2017; Martin et al., 2011). Ma and Chen (2019) provide R code using the package nimble on how to implement a Bayesian semiparametric latent variable model by allowing the latent variable to follow a Dirichlet process prior for joint analysis. They utilize the nimble R package and provide a tutorial of their research. In their dataset, they deal with 10,166 observations (Ma & Chen, 2019). Nimble allows MCMC and Expectation-Maximization (EM) algorithm but comes with a computational burden if the

sample is large (de Valpine et al., 2017). Commonly used for mixed latent variable models in existing literature is also the R package MCMC (Fahrmeir & Raach, 2007; Khatab, 2007; Rosenthal & Voeten, 2007; Samani & Ganjali, 2010). The user specifies the distribution that evaluates the log unnormalized density (Martin et al., 2011).

Fahrmeir and Raach (2007) describe a flexible semiparametric latent variable model allowing for mixed binary, ordinal and continuous indicators. The model consists of a measurement model for direct effects and models indirect effects via a structural model describing the latent variable by covariates. Furthermore, it allows nonlinear spatial effects and covariate effects are modelled semiparametric. For parameter estimation the MCMC method using the R package MCMCpack is applied (Fahrmeir & Raach, 2007). Cai et al. develop a generalized latent variable model for mixed data types using different link function. To conduct their analyses, they use a Bayesian approach and the MCMC method written in R. Their data includes 2,564 observations and the Bayesian estimates take them with a single computer using 2.00 GB RAM 2 hours to complete (Cai et al., 2011).

2.5.2. EM Algorithm

After creating a mixed latent variable model, an estimation algorithm for the distribution can be placed such as Expectation-Maximization (EM) algorithm. Similar to MCMC, EM algorithm use the iteration method until the model converges. The EM algorithm is used to calculate the maximum likelihood estimate using a random sample. It consists of two general steps: In the first E-Step, the expected value for each latent variable is estimated. In the following M-Step the distribution parameters of the models are optimized using a maximum likelihood. The algorithm starts with an initial value for

the model parameters. The EM algorithm does not guarantee the convergence to the global optimum (Amiri et al., 2018; Lin et al., 2016). Furthermore, Lee and Shi state that the EM Algorithm with no missing data is computational faster (Lee & Shi, 2001).

2.6. Common Software Used and Available Packages

2.6.1. Mplus

Muthén is one of the main agents that uses and advocates the Mplus software for his latent variable models (Asparouhov & Muthén, 2016). It is commonly used by other authors for mixed modeling (Depaoli & Clifton, 2015; Lubke & Luningham, 2017; Song et al., 2018; Tillmann et al., 2020; Varriale & Vermunt, 2012; Wall et al., 2012; Yuan et al., 2011). Since Mplus is a commercial software the user community and support is not as large as in the open source software R or even python. To conduct mixture modeling in Mplus the Mixture Add-On is required. Overall, R provides more flexibility with its open-access. The R package MplusAutomation is available and allows an interface between Mplus and R allowing for large-scale analysis but limited in providing e.g. factor loadings from factor analysis models. It allows to run the model in batches (Hallquist & Wiley, 2018).

2.6.2. Statistic Software R

In the more recent years, the free statistic program R has evolved in the mixed modeling area and is frequently used and further developed. Many authors use R especially in the more recent years. (Jafari et al., 2015; Ma & Chen, 2020; Tabrizi et al., 2020; Thmasebinejad & Tabrizi, 2015). The named authors use an optimization package to maximize their objective function which is the likelihood function this approach does

not fit to our research outline. Wendt et al. (2019) use the R lavaan package for their mixture model (Wendt et al., 2019). The lavaan package offers SEM and confirmatory factor analysis functions. Other promising R packages are MCMC or MCMCpack which utilize the BUGS platform for sampling (R Core Team, 2013). A serious advantage for MCMC models is given by the Stan platform which uses Hamiltonian Monte Carlo (HMC) algorithm. In the more recent years, steady interest has been growing for HMC because of its faster computation for complex and large models and available through R packages such as rstan and brms (Monnahan et al., 2017; R Core Team, 2013).

2.6.3. Python and Other Programs

Another language which is a more recent development for mixed models is python. Nussbaum and Giesen developed a python3 package for their mixed model which they made publicly available (Nussbaum & Giesen, 2020). Other authors use Matlab (Kunihama, 2015), Latent GOLD (Morlini, 2012), GLAMM (Rabe-Hesketh et al., 2004), GAUSS programming language (Bhat, 2015), SAS (Yuan & Chan, 2016), SPSS (Ramezani et al., 2019), and BUGS platform (Song & Lee, 2006) for their mixture models.

2.7. Handling Mixed Data Types

The dataset includes 28 relevant variables: 23 binary variables, 2 ordinal variables, 3 continuous variables. This creates two challenges: With 28 potentially relevant variables, there is a high dimensionality. Additionally, we deal with mixed response variables which creates a challenge of how to combine all variables in one model. The proposed plan is to convert the binary variables in the dataset into continuous

data.

In structural equation modeling, the latent variable exhibits a linear relationship with the indicators. This underlines the importance of choosing the right distribution for the indicators. The most common used distribution is the multivariate normal distribution (Lubke & Luningham, 2017). There are two general ways how to include categorical variables into latent variable mixture models: First, a so-called threshold model can be used which treats the categorical variable as continuous. By creating thresholds, it creates partitions of the indicators into response categories. The second way is to replace the linear regression of the categorical variables with either logistic or multinominal regression (Lubke & Luningham, 2017). Another approach is to combine categorical with continuous data is to utilize their correlation: First, polychoric, polyserial and product-moment correlation are computed and then the correlation matrix is used for SEM (Yuan et al., 2011).

The most common way to include binary data in a latent variable model is via the logit or the probit model. The logit model uses the natural logistic distribution link function. Compared to the probit model the logit model has heavier tails and is the most common way in the medical field for including binary variables. The latent variable is predicted in a linear combination of the indicators. In the probit model the standard normal distribution replaces the logistic distribution. From the computational side, the probit model requires more calculation capacity therefore the logit model is typically favored (Khatab, 2007).

2.7.1. Handling Binary Variables: Item Response Theory Model

Since applying a logit model only transfers the binary variable into continuous ones, the idea to summarize the binary variables into a single continuous measure using a dichotomous item response theory (IRT) model provides more benefit in regards of dimensionality reduction. The idea behind IRT models is that behind the data lays a so-called “latent trait” which is linked to the observed variables. This idea is similar to the previous discussed latent variable models such as factor analysis and SEM. Special about IRT models is that they are developed for ordinal or dichotomous data and therefore a good fit as latent variable model for the binary components of the dataset.

In general, IRT models are parametric hence work with a normality assumption for the underlying latent trait. Considering the binary coded complications in birth deliveries we know that most women deliver their babies without these severe complications. This results in a highly skewed dataset, wherefore the normality assumption is violated.

Another assumption for parametric item response theory models is local independence. Local independence means that observed variables are conditionally independent, given the latent variable. It implies that no relationship exists between the variables but the item’s variance is explained by the latent variable (Chen et al., 2013). When the underlying data showcases correlation between the observed variables, a study by Dirlik (2019) proves that nonparametric IRT outperforms parametric IRT models. The parametric IRT is based on a parametric function for the latent trait distribution. In contrast, nonparametric IRT is based on the assumption that the latent trait is only limited by order restrictions (Dirlik, 2019). Based on this knowledge we choose a non-parametric

IRT model.

2.7.2. Skewed Data

Mixture modeling has an explanatory character with its roots in choosing the best model after calculating goodness-of-fit indices. Therefore, it is crucial to be clear on the model's assumptions and take pre-existing knowledge about the model into account (Lubke & Luningham, 2017). Typically, the latent variable underlies the normality assumption such as common in structural equation models and factor analysis models. This assumption is often wrong in real world datasets where normal distributed variables are rather rare (Yuan, 2016).

The normality assumption does not apply to the dataset, therefore it creates a third challenge: the skewness of the latent factor. Given that the dataset shows most birth deliveries have no complications more mothers and newborns will be healthy and therefore create a skewed latent variable. Considering non-normal data, Azzalini and Valle (Azzalini & Valle, 1996) have laid the foundation for further research by addressing issues with non-normal data when applying models with a normality assumption and proposing the multivariate-skew distribution. There has been multiple studies in data models for continuous data in the recent years taking skew-normal, student-t and mixtures of skew-normal distributions in account (Asparouhov & Muthén, 2016; Contreras-Reyes & Arellano-Valle, 2013; Lin et al., 2014).

The most common way to handle skewed data is to standardize and transform the data to approximate the distribution to a normal distribution. For applying data transformation, the log transformation is the most commonly used method (Azzalini & Valle, 1996; Rezaei Ghahroodi et al., 2019). It is used by various researchers in order to

ensure the normality assumption of their latent variable models (Amiri et al., 2018; de Leon & Carrière, 2005; Feng et al., 2019; Kunihama, 2015; Nussbaum & Giesen, 2020; Quinn, 2004; Samani & Ganjali, 2010; Tillmann et al., 2020).

Another way to deal with skewed variables in a latent variable model is to replace the normal distribution. Without using the log transformation, we can directly estimate the hidden variable by using different distributions. De Leon and Cough describe in their book Analysis of Mixed Data Chapter 11 an alternative to the commonly used log normal transformation. In their example they deal with a highly positively skewed variable and where they use a model utilizing the gamma distribution (de Leon & Chough, 2013). This approach requires previous knowledge about the latent variable distribution.

2.8. Difference to Existing Literature

Meijer et al. (2011) introduce an approach to address the lack of a quantitative measure for health status. Since health cannot be measured directly but is a result of different economic, physical, and social indicators, they establish an internationally comparable health index by using a latent variable model. For their research, they use self-reported mobility surveys from Europe. Meijer et al.'s publication where they use 29,835 observations to estimate a latent health variable using mixed data types, which is the biggest sample size we found in the literature. In the end, their health index reaches a reliability of 80% (Meijer et al., 2011). This study is different from theirs, as this study focuses on quantifying health status regarding birth delivery.

Latent variable models – using mixed response variables as indicators and a continuous latent factor – has been intensively studied in the recent years. Bhat proposes a generalized heterogenous data model for mixed data indicators. It utilizes the maximum

approximate composite marginal likelihood for model estimation, which reduces computation time for high dimensionality models. The model is applied on a simulated dataset with 57 parameters to estimate and a maximum sample size of 3000 (Bhat, 2015). A different approach to handle mixed data for latent variables is proposed by Yuan. He summarizes current meta-analytical SEM models and proposes procedures for dealing with nonnormality. He emphasizes robust methods and adjusted test statistics to account for nonnormality in SEM models. Furthermore, he developed a ridge function which can be applied to the maximum likelihood or generalized least squares estimation. This yields in more accurate results for skewed datasets with heavy tails (Yuan, 2016). While their methods are useful for our research, both authors work with dataset from different fields. Bhat (2015) uses transportation data and Yuan (2016) works with education data.

We use the Vital Birth Record 2018 data, which contains 28 relevant variables and 3,801,534 observations prior to the data cleaning. However, in existing literature big data sets for mixed models (Meijer et al., 2011) are only a fraction compared to the Vital Birth Record 2018 dataset with roughly 3.8 million observations and 28 relevant mixed data type indicators. This creates another challenge of the proposed thesis: some of the proposed methods may not apply with such a huge dataset as the Vital Birth Records 2018.

In conclusion, the present Vital Birth Record 2018 dataset exhibits very unique features characterized by extremely skewed variables due to rare complications, high dimensionality, a huge sample size, and the mixed data types. We conclude that the hidden variable will be highly abnormal, so existing software packages cannot apply directly.

2.9. Performance Measures

Standardized measures are needed to assess how precise and reliable a model actually is. Models of different types, namely factor analysis, item response theory, cluster analysis, and MCMC have their own model performance measures which makes it hard to directly compare different model types. Furthermore, new performance measures are constantly developed which may create a confusion about which indices to use (Rex, 2016). Hereby, we provide an overview about common performance measures, including goodness-of-fit indices, residual measures and several model-specific performance measures.

2.9.1. Hypothesis Testing

SEM tests the multivariate relationship between observed and latent variables. It incorporates confirmatory factor analysis and path analysis in one model. To evaluate the SEM model, hypothesis testing is used to assess the model-fit. The hypothesis tests the discrepancy between model covariance matrix and original covariance matrix. The null hypothesis states a good fit between the researcher's model and the observed data. The alternative hypothesis is the null model – also called baseline model – which assumes no correlation between the numerical variables and only includes mean and variance of each observed variable. If the model is specified correctly, the null hypothesis would *not* be rejected and the p-value would be insignificant (Xia & Yang, 2019). The test statistic is calculated as follows:

$$\text{Chi-square test statistic} = (n-1) * f,$$

where f is a value of discrepancy between the data and the model produced

covariance matrices and n is the sample size. It is estimated using the model's parameters and maximum likelihood and therefore assumes a multinormal distribution of the underlying joint population distribution of the observed variables (Kenny, 2020; Rex, 2016).

An optimal SEM model would show no significant difference between model and original covariance matrix. Thus, an optimal SEM model would show a p-value larger than the significance level resulted from a chi-square statistic close to zero (Fan et al., 2016); in other words, the null hypothesis would not be rejected and the p-value would be insignificant (Xia & Yang, 2019).

Rex recommends in his handbook for Principles and Practice of Structural Equation Modeling to report the model's chi-square with its degrees of freedom and p-value. More degrees of freedom automatically reduce the value of the chi-square test statistic which would lead to overparameterization and drastically bias the model. The hypothesis test is further influenced by correlation among the observed variables of the model. The chi-square test statistic is directly influenced by the sample size: a large sample size causes the model to get overly sensitive to small data anomalies. Therefore, the analysis of residuals is more important for large sample sizes than merely looking at the test statistics (Rex, 2016).

2.9.2. Goodness-Of-Fit Indices

The **Comparative Fit index (CFI)** is a measure which compares the model versus the so-called baseline model or null model. The null model assumes no correlation between the observed variables. The CFI is calculated as follows:

$$\frac{\chi 2 - df\ (\text{Null Model}) - \chi 2 - df\ (\text{Proposed Model})}{\chi 2 - df\ (\text{Null Model})}$$

The fit function uses the polychoric correlation matrix of the specified model as base for calculation. For scaled statistics a scaling or shifting parameter is added to the calculation (Xia & Yang, 2019). Its value lays between 0 and 1 and describes the relative value between the baseline model and the research model. For example, a CFI of 0.6 shows that the model is 60% stronger than the baseline model which is the weakest possible model. Therefore, a CFI close to 1 implies that the model fits the data well (Kenny, 2020; Rex, 2016).

Tucker-Lewis Index (TLI) – also called non-normed fit index (NNFI) – measures the model fit in a similar manner as the CFI. Though TLI also calculates the difference between the baseline model and the test model; the baseline model is considered worst fit, which assumes uncorrelated observed variables The TLI is calculated as follows:

$$\frac{\chi 2/\text{df}(\text{Null Model}) - \chi 2/\text{df}(\text{Proposed Model})}{\chi 2/\text{df}(\text{Null Model}) - 1}$$

The TLI takes values between 0 and 1. Though TLI is technically not restricted on the upper end, any value larger than one come to be restricted to one. A larger index shows a better model fit, for example, a TLI of 0.9 shows that the proposed model is 90% better than the null model. Compared to the CFI, the TLI is more sensitive to the sample size. However, the CFI applies a penalty for added parameters and thus is more commonly reported. Since TLI and CFI are correlated, only one of these two measures needs be reported for SEM or factor analysis model evaluation (Kenny, 2020; Rex,

2016).

Regression models assess model fit using the **R^2 metric** which provides the amount of variance explained by the model. This metric does not apply to Bayesian models directly. To incorporate a valid measure between 0 and 1 for Bayesian models Gelman et. al. proposed the Bayesian R2 and the Loo R2 value which are provided for MCMC outcomes.

The **Bayesian R^2** measure is comparable to the R^2 value of a linear regression while the Loo R^2 is comparable to an adjusted R^2 value for linear regression.

LOO R^2 calculates the R^2 statistic adjusted by LOO residuals which incorporate the unknown posterior data distribution. The basic difference between the Bayesian R^2 and LOO R^2 is the input data for its calculation: the Bayesian R^2 score is calculated by using the given dataset, but the Loo R^2 algorithm creates new data based on the dataset's frequencies and hence uses "independent" data (Gelman et al., 2019). LOO R^2 uses the LOO residuals while the Bayesian R^2 uses the modeled residual variance. For the detailed calculation see Gelman, Goodrich et al. (2019).

In general, goodness of fit indices should serve as an approximation method to an acceptable or unacceptable model rather than being judged without the researcher's theoretical knowledge about the model specification (Rex, 2016).

2.9.3. Predictive accuracy

Akaike Information Criterion (AIC) is a popular model fit measure which describes a model's fitness to the data. It is calculated as follows:

$$AIC = -2\ln(L) + 2k,$$

where k is the number of model variables and L is the model's maximum likelihood

estimation. The reason for its popularity yields in its power to combine model selection and statistical estimation methods. It allows to compare different model types based on the same dataset. The AIC is low when the likelihood estimation fits the data well. Hence, a good model would have an AIC close to zero (Zajic, 2019).

Bayes Information Criterion (BIC) is a similar measure as the AIC, it also uses the model's maximum likelihood to estimate the model's fit to the data. The BIC is calculated as follows:

$$BIC = -2\ln(L) + k \log n,$$

where k is the number of model variables, n is the sample size, and L is the model's maximum likelihood estimation. Compared to the AIC, the BIC includes the model's sample size in its calculation. Compared to the BIC, the AIC includes a greater penalty for the number of parameters. To compare different models, both AIC and BIC should be provided if possible (Fabozzi et al., 2014). A better model fit applies when both measures, AIC and BIC, are smaller. However, information criteria may be subject to sampling error. Both AIC and BIC can increase with larger sample sizes. Therefore, researchers should be cautious when selecting a model based on information criteria (Rex, 2016).

The **Watanabe-Akaike Information Criterion (WAIC)** is a widely applied information criteria specific in MCMC models. Compared to the AIC, WAIC averages over the posterior distribution instead of the point estimates. It uses the maximum likelihood from the underlying Bayesian model to calculate pointwise prediction accuracy out of the given sample. It is scaled to allow direct comparison with the AIC and defined by (Watanabe, 2010) as

$$\mathrm{WAIC}(n) \equiv B_t L(n) + \frac{\beta}{n} V(n),$$

where $B_t L(n)$ is Bayes training loss, V(n) is the functional variance and $0 < \beta < \infty$. Smaller values indicate a better model-fit.

To assess the reliability of WAIC in a MCMC model, we need to look at the indicator **Pareto k estimates**. The Pareto k estimates monitor each sampling step and provides an overview of influence of each observation on the posterior distribution. The Pareto k estimates range between 0 and 1. A threshold of 0.7 is commonly used to identify highly influential observations which distort convergence; that is, Pareto k values above 0.7 imply unreliable MCMC estimates. Thus, additional computations to increase the reliability are necessary and the model should be altered to a more robust form. In general, the WAIC should not be trusted if Pareto k estimates are detected above 0.5 (Paananen et al., 2021; Vehtari et al., 2017).

Leave-one-out cross validation (LOOIC) is a measure for the model's predictive accuracy. A LOOIC value close to zero indicates a good model fit. While the AIC calculation assumes multivariate normality and does not take prior assumptions in account, the LOOIC does not assume any distribution and incorporates the uncertainty of parameters. Therefore, it is a measure commonly used in MCMC models (Vehtari et al., 2016). It is calculated as follows using importance sampling:

LOOIC = -2*elpd_loo,

where elpd_loo is the expected log pointwise predictive density for the Bayesian leave-one-out estimate. Compared to the WAIC for MCMC model fit evaluation, the LOO is more robust and works better when having weak priors or influential observations

(Vehtari et al., 2016). Even though WAIC and LOOIC are more advantageous in MCMC, compared to the more common used information criterion AIC, they require a longer computation time. This is why they are unusual in practical usage (Vehtari et al., 2016).

2.9.4. Residual Analysis

Root Mean Square Error of Approximation (RMSEA) measures the absolute deviance from the approximated model between the hypothesized model and the covariance matrix of the underlying population. It ranges between 0 and 1, where a good model should have a RMSEA close to zero (Hoyle, 2011). Threshold parameters for the RMSEA value are controversial, some researchers advise that a RMSEA value above 0.1 might imply a crucial model problem (Rex, 2016). It is calculated as follows:

$$RMSEA = \sqrt{\frac{\chi^2}{\frac{df}{N-1}} - 1}$$

Standardized Root Mean Square Residual (SRMR) calculates the absolute standardized difference between observed and predicted correlation. It represents a standardized version of the general root mean square residual which is heavily dependent of the variable's metrics. The SRMR applied a standardization term in order to make the residual interpretation easier. It is calculated as follows:

$$SRMR = \sqrt{\frac{1}{2}\sum (S_{ij} - I_{ij})^2} \quad ,$$

where S is the sample correlation matrix and I the implied correlation matrix (Prudon, 2015). A SRMR value closer to zero indicates a better model fit. Therefore, the

researcher should be careful when detecting SRMR values above 0.1 (Kenny, 2020; Rex, 2016).

Root Mean Squared Error (RSME) describes the standard deviation of the model residuals and tells the researcher how the residuals deviate on average from the model prediction. Its unit is dependent of the scale for the measured variable. A better model fit would showcase a smaller RSME. It is calculated as follows:

$$\text{RMSE} = \sqrt{1 - r^2}\ SD_y,$$

where SD_y stands for the standard deviation of the predicted y value, and r are the residuals of the model. The idea is to square residuals of a model, take their average, and then calculate the square root of the result (Glen, 2021c).

2.9.5. Model-Specific Indices

The **modification index** provides guidance on how to improve an existing SEM model. In SEM models, the latent variable is estimated by a set of linear correlated variables. However, other model paths, such as correlation between latent variables and residual correlation of the observed variables, are often unnoticed. To identify which further model structures are important for the model specification, the modification index evaluates different model paths which have not yet been specified in the SEM model. A path is for example a residual correlation or the connection between observed variables and the latent variable. In a SEM model, every path which has not been included in the model specification is fixed to zero. This creates so-called fixed paths whereas the model parameters are freely estimated. The modification index evaluates which of those fixed parameters of the model should be specified and thus freely estimated. It analyzes what-if scenarios by changing fixed parameters to freely estimated parameters and calculates the

model chi-square improvement if a certain path would be specified by the user. A large modification index tells that the corresponding path is valuable to specify and allow to be freely estimated.

Significant path structures which are not yet included in the model are signaled by a high modification index. The general rule is to include paths with the highest modification index. By including a path pointed out by a high modification index, the model loses one degree of freedom, but its chi-square improves by the estimated modification value for the corresponding path. The modification index is calculated by a general fit function based of the maximum likelihood. (Rex, 2016; Sörbom, 1989). For further calculation definition see Sörbom (1989).

In factor analysis and SEMs, the factor loading indicates how much of the latent variable is explained by this factor, where factor refers to variable. How large a factor loading should be to be considered important or useful is controversial among scientists. Some researchers propose that a factor loading must be above 0.5 to be considered important for the model, while others argue that a factor loading of 0.4 or higher can be useful (Meyer, 2020). For explanatory factor analysis the **sum of squared (SS) loadings** are another measure which evaluates the usefulness of a factor. The rule is to keep SS loadings above 1 since they are considered useful for the model (Ford, 2016).

Communality – denoted as h^2 – describes a predictor variable's usefulness in factor analysis models. It measures how much of the common variance in the dataset is explained in a particular observed variable and ranges from 0 to 1. Common variance – also called shared variance – explains a variable's variance shared with other variables given the common factors (Meloun & Militky, 2011). If a variable is independent from

other variables in the dataset, its communality would be zero. For example, if a variable's communality is 0.7, this means 70% of its variance follows the other variables' variance in the dataset, and can be described by a latent variable which explains the shared variance. Looking at a particular observed variable, a communality close to one shows that the model explains almost 100% its variance (Glen, 2021a).

3: DATA PREPARATION

3.1. Data Cleaning

As a first step we clean the dataset. Since the dataset is very big and we have less than 1% missing data for each variable, observations with missing values are deleted. While the dataset shall only contain live births, we find 758 cases of stillbirths which we exclude for further calculations. After cleaning the data, we continue with 3,769,386 out of 3,801,534 observations (0.846% missing data).

3.2. Preprocessing Binary Variables

The dataset contains 23 binary variables describing maternal morbidity, abnormal conditions and genetic disabilities of the newborn. The variables are recorded binary, yes if occurred, no if not occurred and the frequency is shown below in Figure 3. Most US births are delivered healthy. Therefore, we observe only small fractions of severe complications in these categories. The largest occurrence is *Admission to Neonatal Intensive Care Unit* which has a total occurrence of 9.03% in 2018. The smallest category is *Anencephaly,* a serious birth defect with parts of the brain and skull missing (Center for Disease Control and Prevention, 2020a), with a total occurrence of 309 cases out of 3,769,386 birth deliveries in 2018. As a second step, we will focus on solving the high-dimensionality issue in the next subsection.

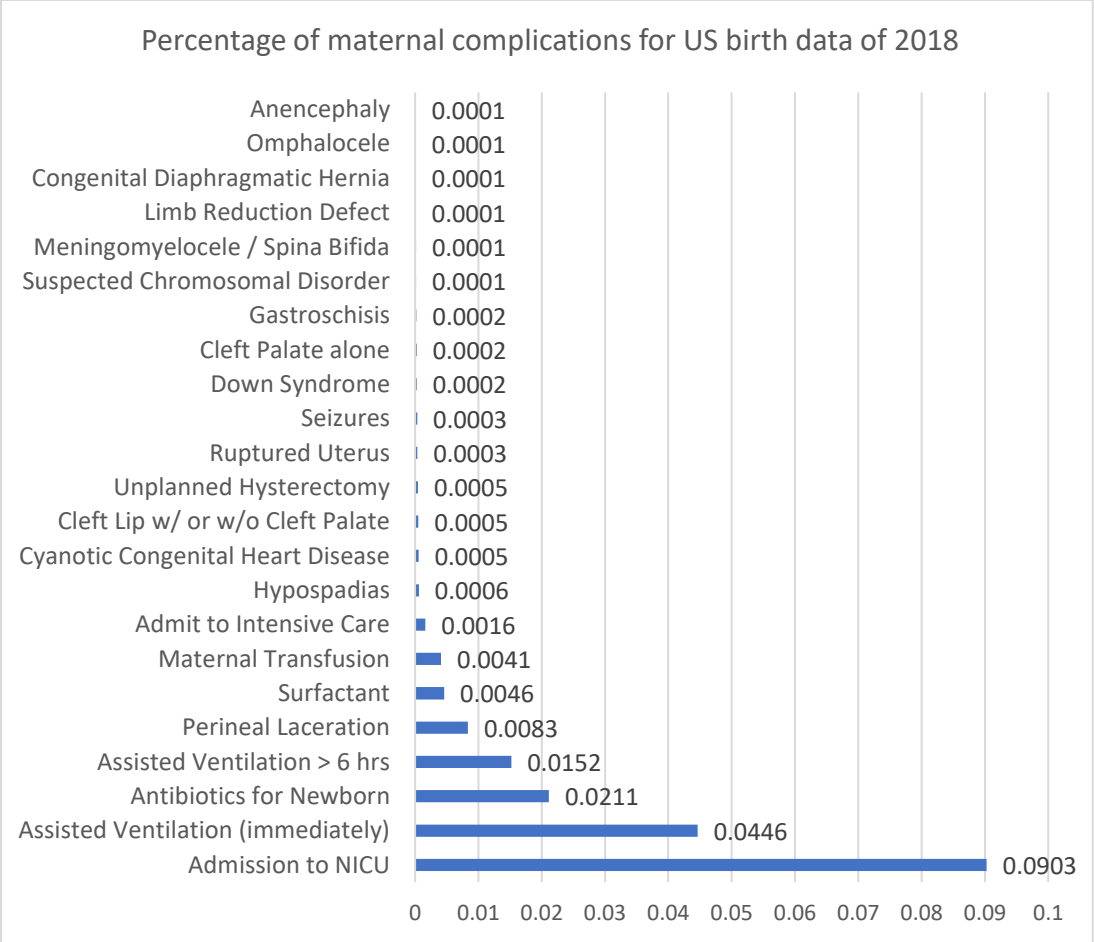

Figure 3: Frequency chart of binary variables for US births 2018

3.2.1. Feature Selection

To address the challenge of high dimensionality, we perform feature selection. In total, two different methods for feature selection were applied using python (Jupyter Notebook). The dataset was too large to perform a lap_score package from skfeature in python. In literature for latent variables, authors applied the LASSO method for feature selection which is described as "Bayesian least absolute shrinkage and selection operator" (Wang et al., 2020). It can be used on ordinal variables but comes with certain limitations: a cutoff parameter must be specified beforehand determining the importance of a covariate. It also requires a standardization of the variables on the same scale (Feng

et al., 2017; Wang et al., 2020). Since we cannot validate the feature selection outcome as the target variable is latent, we apply multiple feature selection methods and select the common variables found relevant in all feature selection methods: low variance, principal component analysis, and mixed principal component analysis.

3.2.1.1. Low Variance Method

The first applied feature selection method is the low_variance method from the skfeature package (Li et al., 2017). The method required a variance threshold as cutoff parameter for the variable selection. Variables with high variance are selected based on the indicated cutoff parameter. In total, four different cutoff parameters were specified [0.001, 0.0005, 0.0003, 0.0001] resulting in 8, 11, 14, and 21 selected variables respectively. Results are presented below:

Table 2: Feature selection using low variance method

		threshold parameter			
Category	**Variable name**	**0.001**	**0.0005**	**0.0003**	**0.0001**
Maternal Morbidity	Maternal Transfusion	x	x	x	x
	Perineal Laceration	x	x	x	x
	Ruptured Uterus			x	x
	Unplanned Hysterectomy			x	x
	Admit to Intensive Care	x	x	x	x
Abnormal Conditions of the newborn	Assisted Ventilation (immediately)	x	x	x	x
	Assisted Ventilation > 6 hrs	x	x	x	x
	Admission to NICU	x	x	x	x
	Surfactant	x	x	x	x
	Antibiotics for Newborn	x	x	x	x
	Seizures			x	x
Congenital abnormalities	Anencephaly				
	Meningomyelocele/ Spina Bifida				x
	Cyanotic Congenital Heart Disease		x	x	x
	Congenital Diaphragmatic Hernia				x
	Omphalocele				
	Gastroschisis				x

		threshold parameter			
Category	**Variable name**	**0.001**	**0.0005**	**0.0003**	**0.0001**
Congenital abnormalities	Limb Reduction Defect				x
	Cleft Lip w/ or w/o Cleft Palate		x	x	x
	Cleft Palate alone				x
	Down Syndrome				x
	Suspected Chromosomal Disorder				x
	Hypospadias		x	x	x

By comparing the results from the low variance method in Table 2 with the frequency of occurrence in Figure 3, we notice an association between a high frequency of occurrence (Figure 3) and the selection of this variable in Table 2. This can be explained by the algorithm of the low variance method: it selects variables with a high variance, so we notice that complications with higher frequencies of occurrence (Figure 3) exhibit a higher variance and are selected in all cutoff threshold.

3.2.1.2. Principal Component Analysis (PCA)

The second method has been applied to the dataset: principal component analysis (PCA). This method is not considered a feature selection method, but it identifies so-called principal components which are describe of artificial variables constructed using linear mixtures of the original variables from the dataset. The advantage is to reduce dimensionality by trying to lose as the least amount of information as possible. Additionally, correlated variables are merged and just the most important features which explain the highest variance are extracted from the dataset (Malik, 2018). PCA is conducted on the binary variables coded with 0 and 1, where 0 stands for not occurred and 1 for event happened. The number of components can be either visualized in a Scree-plot or calculated. Before determining the number of components, we need to

decide how much variance should be explained by the principal components. Typically, an explained variance of 95 – 99% is chosen (Mikulski, 2019). Since we are in an explanatory stage, we try both options. The calculation results in seven principal components which explain 95% of the variance in the binary variables and thirteen principal components explaining 99% of the variance and the result of the PCA is provided in the appendix. With only 4% difference between these two options, the number of variables selected, the 99% method almost doubles. A summary of the selected variables is presented in Table 4.

3.2.1.3. Mixed PCA

The third method to reduce dimensionality is the mixed Principal Component Analysis (mPCA). It utilizes the same vectorization methods to summarize information as the general PCA but allows categorical and numerical data types in the same model. Specifically, mPCA allows to evaluate the importance of each variable through their impact of the vector loadings (Principal Components) with all considered numeric and binary variables. Since PCA and factor analysis share common ideas, the factor loading cutoff 0.4 is used for the mPCA in order to identify variables with significant impact (Meyer, 2020; Santos et al., 2019). When loosening the criteria, we also consider slightly weaker vector loadings of 0.3 or higher for our model. The resulting squared loading matrix of the mPCA is provided in the appendix for further detail. The summary of selected variables by mPCA is presented in Table 3.

Table 3: Selected features based on mPCA method

	Categorized as	cut at 0.3 loading*	cut at 0.4 loading*
Numeric		Apgar 5	
		Apgar 10	Apgar 10
		Mother's age	
		Birth weight	Birth weight
		Combined gestation	
Binary	Maternal Morbidity	Maternal transfusion	Maternal transfusion
		Perineal Laceration	Perineal Laceration
		Ruptured Uterus	Ruptured Uterus
		Unplanned Hysterectomy	Unplanned Hysterectomy
		Admit to Intensive Care	Admit to Intensive Care
	Abnormal conditions of the newborn	Assisted Ventilation (immediately)	Assisted Ventilation (immediately)
		Assisted Ventilation > 6 hrs	Assisted Ventilation > 6 hrs
		Admission to NICU	Admission to NICU
		Surfactant	Surfactant
		Antibiotics for Newborn	
	Congenital abnormalities	Anencephaly	Anencephaly
		Cyanotic Congenital Heart Disease	Cyanotic Congenital Heart Disease
		Congenital Diaphragmatic Hernia	
		Gastroschisis	
		Limb Reduction Defect	
		Cleft Palate alone	Cleft Palate alone
		Cleft Lip w/ or w/o Cleft Palate	Cleft Lip w/ or w/o Cleft Palate
		Suspected Chromosomal Disorder	
		Hypospadias	
Total number of variables			
N		24	18

*vector loadings rounded down (e.g. 0.35 is 0.3)

Since the mPCA method is the only method which allows us to evaluate the numeric variables together with binary ones, it allows us to evaluate the relevance of the numeric variables: the two for the numeric variables with the highest vector loading are birth weight and Apgar 10. They both load into the third and fourth principal component and show a vector loading above 0.4. When we loosen the criteria to a loading cutoff of 0.3, the principal components which explains most of the variance is constructed by Apgar 5 and the mother's combined gestation time, assisted ventilation immediately and >6 hours, and admission to the neonatal intensive care unit. The second dimension is

mainly described by the binary complications: maternal transfusion, unplanned hysterectomy, and the mother's admission to the intensive care unit. The third principal component mainly consists of both Apgar scores, while the fourth principal component is characterized by the baby's birth weight and the mother's combined gestation time. Please note that the mother's age does not play a significant role in any of the principal components, as it shows up with a total loading of 0.358 in the ninth principal component. This indicates that the association of mother's age with birth delivery quality may not be close, compared to other numerical variables. In total, the mPCA with 99% explained variance reduces our variables to 27, whereas the last two dimensions do not exhibit any relevant factor loadings using a cutoff value of 0.3.

3.2.1.4. Feature Selection Summary

In Table 4, we provide an overview of all applied feature selection methods and the specified cutoff values.

Table 4: Summary of all feature selection method's cutoffs

	low variance				**PCA***		**99% variance mixed PCA****		**total**
criteria	0.001	0.0005	0.0003	0.0001	95%	99%	cut at 0.4	cut at 0.3	
Maternal Transfusion	X	X	X	X	X	X	X	X	8
Perineal Laceration	X	X	X	X	X	X	X	X	6
Ruptured Uterus			X	X		X	X	X	3
Unplanned Hysterectomy			X	X		X	X	X	5
Admit to Intensive Care	X	X	X	X		X	X	X	7
Assisted Ventilation (immediately)	X	X	X	X	X	X	X	X	8
Assisted Ventilation > 6 hrs	X	X	X	X	X	X	X	X	8
Admission to NICU	X	X	X	X	X	X	X	X	8
Surfactant	X	X	X	X	X	X		X	7

	low variance				PCA*		99% variance mixed PCA**		total
criteria	0.001	0.0005	0.0003	0.0001	95%	99%	cut at 0.4	cut at 0.3	
Antibiotics for Newborn	X	X	X	X	X	X		X	7
Seizures			X	X					2
Anencephaly							X	X	0
Meningomyelocele / Spina Bifida				X					1
Cyanotic Congenital Heart Disease		X	X	X		X	X	X	4
Congenital Diaphragmatic Hernia				X				X	1
Omphalocele									0
Gastroschisis				X				X	3
Limb Reduction Defect				X				X	1
Cleft Lip w/ or w/o Cleft Palate		X	X	X		X	X	X	6
Cleft Palate alone				X			X	X	3
Down Syndrome				X					1
Suspected Chromosomal Disorder				X				X	2
Hypospadias		X	X	X		X		X	4
number of features selected	**8**	**11**	**14**	**21**	**7**	**13**	**12**	**19**	

*relevant loadings > 0.7

** criteria: cutoff for relevant loadings

This leads us to the following selected variables by each method, presented in Table 5.

Table 5: Summary of selected features by all feature selection methods

Categorized as	Selected by all methods	Selected by variance 0.0003, sklearn 99% and cut at 0.3
Maternal Morbidity	Maternal Transfusion	Maternal Transfusion
	Perineal Laceration	Perineal Laceration
		Ruptured Uterus
		Unplanned Hysterectomy
		Admit to Intensive Care
Abnormal conditions of the newborn	Assisted Ventilation (immediately)	Assisted Ventilation (immediately)
	Assisted Ventilation > 6 hrs	Assisted Ventilation > 6 hrs
	Admission to NICU	Admission to NICU

Categorized as	Selected by all methods	Selected by variance 0.0003, sklearn 99% and cut at 0.3
		Surfactant
		Antibiotics for Newborn
Congenital abnormalities		Cyanotic Congenital Heart Disease
		Cleft Lip w/ or w/o Cleft Palate
		Hypospadias
	Total	**Total**
	5	**13**

Similar to the feature selection by threshold variance cutoff, we observe common variables which are of high relevance in the PCA as well. In the category *Maternal Morbidity* we observe *Maternal Transfusion* and *Perineal Laceration* as important characteristics in all three dimension reduction methods. When relaxing the cutoff in all methods, we additionally observe *Ruptured Uterus, Unplanned Hysterectomy* and the *Admit to Intensive Care* as relevant in the Maternal Morbidity category. The variable *Admit to Intensive Care* seems relevant in the variance cutoff method, but only loads with a high weight in the eighth component in the 99% explained variance and contributes 0.008 to the overall explained variance. The two other variables in this category, *Ruptured Uterus* and *Unplanned Hysterectomy*, explain each with 0.002 the least of the 99% variance and do not appear in the first two thresholds (0.001, 0.0005) in the variance cutoff method either. However, both variables are relevant in the mixedPCA method with 99% variance explained, where they are both load more than 0.4 into the principal components.

The results from the feature selection match the clinical insights. Maternal transfusion describes the condition during birth delivery where a blood transfusion needs to be given to the mother, if she bleeds too much during birth delivery. Furthermore, maternal transfusion is seen as one of the major indicators of maternal morbidity, since it

is a common occurrence in cases of severe maternal morbidity (Center for Disease Control and Prevention, 2020b). Relevance of maternal morbidity is also captured between the third or fourth degree perineal laceration, which describes a tear of a anal muscle layer controlling bladder and bowel functions. Experiencing a third or fourth degree of the muscle tear can lead to a persistent incontinence that requires further medical attention and hospitalization. This complication typically requires surgery and at a minimum several weeks of medical care (Mayo Clinic, 2019). The third or fourth degree tear also impacts future birth deliveries because women are 4-times more likely to repeat the complication in later birth deliveries than those without this complication (Woolner et al., 2019).

A ruptured uterus is a severe complication which may happen during a natural birth. The uterus tears and causes the baby to glide into the mother's abdomen, which leads to life-threatening bleeding and may strangulate the baby. Furthermore, it can cause an unplanned hysterectomy and about 6% of all unborn babies cannot survive this complication. Fortunately it is a rare condition which happens in less than 1% of birth deliveries (Cirino, 2017).

Unplanned hysterectomy describes an emergency surgery in which the uterus of the mother needs to be removed to increase survival chances of the mother. This procedure is due to different reasons, such as uterine rupture, placenta accreta (the placenta grows too deep into the uterine wall), uterine atony (no contractions in the uterus after birth delivery), and life-threatening bleeding. Most commonly it affects women with a history of cesarean section. Unplanned hysterectomy is a severe complication and can lead to about 23% potential death of both mother and child. If the

mother survives, she may suffer physically from bladder injuries, wound infection, and fever; furthermore she is unable to carry any more children (Machado, 2011). Psychologically, the long-term consequences are severe emotional distress, and trauma resulting in a lower life-quality as reported by survivors (Elmir et al.). In general, an unplanned hysterectomy happens in 0.24 to 8.7 cases per 1000 birth deliveries (Machado, 2011).

In the category *Abnormal Conditions of the Newborn,* we observe the variables *Assisted Ventilation (immediately)*, *Assisted Ventilation >6 hours,* and *Admission to Neonatal Intensive Care Unit (NICU)* in all feature selection methods. *Surfactant,* and *Antibiotics for Newborn* are considered significant contributors to the birth delivery quality, especially when the cutoff criteria in all methods becomes more lenient. In the PCA, the variable *Admission to NICU* is the main loading of the first principal component which explains the most variance (0.56). *Assisted Ventilation (immediately)* is the main component of the second principal component and explains 0.185 of the variance. *Antibiotics for Newborn* is with 0.923 the main loading for the third principal component which explains 0.09 of the data's variance. *Assisted Ventilation >6 hours* is with -0.767 the major component of the fourth principal component, which only explains 0.046 of the data's variance. Even less important is *Surfactant* compared to the other items mentioned before; it describes the main loading of the seventh and therefore last component, which only explains 0.020 of the 95% variance PCA. Of less importance is the variable *Seizures* in this category in all dimensionality reduction methods, it describes a moderate weight of the last principal component for explaining 99% variance where the principal component explains 0.002 of the variance.

Admission to the NICU is considered the major indicator for the quality of birth

delivery based on the feature selection methods. From a clinical standpoint, this is an important indicator of the baby's well-being as well. There are many different reasons why admission to NICU is necessary, such as seizures, birth defects, respiratory issues, low birthweight, low height, and a pre-term birth delivery (Stanford Children's Health, 2021). Other conditions which account for the baby's health are both information for assisted ventilation: assisted ventilation required immediately following delivery, and assisted ventilation required for more than six hours. While mechanical ventilation is necessary to save the newborn's life, it also leads to ventilator-caused injury which is associated with severe morbidity and mortality in newborns (Chakkarapani et al., 2020). Besides respiratory failure, a deficiency of pulmonary surfactant is a major contributor for the newborn's health. A newborn is given surfactant replacement therapy if respiratory distress occurs, and the lung alveoli are instable. It is administered in severe respiratory failure cases, such as pneumonia and sepsis (Melbourne, 2018). Another major indicator for the baby's health is described with antibiotics received by the newborn for suspected neonatal sepsis. This is also reflected in the literature and clinical guidelines since antibiotics are a first response treatment for neonatal sepsis. In general, neonatal sepsis is the third major cause of neonatal death and can cause severe disability if survived (Korang et al., 2019).

In the last binary complication category and *Omphalocele* is considered not relevant at all in any feature selection method. In the variance threshold method, *Cyanotic Congenital Heart Disease, Cleft Lip w/ or w/o Cleft Palate* and *Hypospadias* are considered relevant using a threshold parameter of 0.0005 or lower. However, they are not relevant using a threshold parameter of 0.001. In PCA, the three variables appear in the latter principal components 9,10, and 11 and each explains 0.003 of the variance. In mPCA they have a vector loading of 0.3 but below 0.4. *Anencephaly* seems to be only

relevant in mPCA where it appears as a relevant complication in all loading cutoffs. However, *Anencephaly* does not get selected as relevant in any other feature selection method. The remaining eight complication variables, *Congenital abnormalities*, *Meningomyelocele / Spina Bifida*, *Congenital Diaphragmatic Hernia*, *Gastroschisis*, *Limb Reduction Defect*, *Cleft Palate alone*, *Down Syndrome*, and *Suspected Chromosomal Disorder* are less relevant for birth delivery quality in all dimensionality reduction methods. They only appear relevant in the lowest threshold parameter category (0.0001) of the variance cutoff method and do not play any significant role in the principal components either.

Cyanotic Congenital Heart Disease describes a common heart defect of the newborn in which the heart structure did not grow correctly. It usually requires surgery. If left untreated, it would result in severe heart disease complications, such as stroke and heart failure, and is a common cause of death. Babies with Cyanotic Congenital Heart Disease have a 1-year survival chance of 75% and an 18-year survival chance of 69%, and it causes about a third of all infant deaths. It occurs more likely in second pregnancies and has a total occurrence of 8 to 9 in 1000 birth deliveries (Ossa Galvis et al., 2021).

A cleft lift describes an opening of the baby's upper lip. The split can be small or range from the inside of the nose to the upper lip. A split of the upper inner mouth is called cleft palate. It usually causes major problems with food intake and speaking but is also associated with ear infections and teeth problems. It requires initial surgery in the first 12 months after birth delivery and a majority of children needs additional surgery later in their lives as well. The incidence of a *cleft lip with cleft palate* is 1 in 1,700 births,

a cleft lip without cleft palate occurs once in every 2,800 births (Centers for Disease Control and Prevention, 2020).

Hypospadias is a congenital condition in which the urinal tube of a boy's penis is on the underside of the penis instead of its tip. It requires surgery and usually has long-term effects on the newborns urination and ejaculation later in life (Mayo Clinic, n.d.).

3.2.2. Binary Variable Transformation: Item Response Model

In this section, we first compare parametric and non-parametric IRT models using all 23 binary variables. Then we determine the best set of binary variables for IRT estimation of the continuous latent trait.

3.2.2.1. Parametric and Nonparametric IRT Model Performance

The IRT model has been chosen to represent the multiple binary variables using one continuous latent trait variable. This latent trait represents a score of the overall complications occurring during birth delivery and is used as a continuous variable for the upcoming factor analysis and SEMs.

From the frequency table for the binary variables shown in data description section 3.2., we know that all binary variables have a rare occurrence. The binary variable of the most frequent occurrence is Admission to NICU with about 9%. Thus, these binary variables are highly unbalanced. Most IRT models rely on a parametric approach and assume an underlying normal distribution of the latent trait. Since this is not always true in realty, approaches for non-parametric item response model has been developed (Reise et al., 2018).

We use the R package sirt that offers a non-parametric estimation for Rasch-type

models. This package fits our goal to estimate a continuous latent trait with unbalanced dichotomous responses (R Core Team, 2013). Furthermore, the sirt package offers a log-linear kernel distribution smoothing for three moments respectively. Kernel-based smoothing methods are the most common way to establish a non-parametric estimator (Xu & von Davier, 2008).

To see if a three-moment log-linear kernel distribution smoothing improves the non-parametric IRT model, we compare a non-parametric model with three moment kernel-smoothing terms to one without those kernel-smoothing terms. Additionally, the two non-parametric IRT models have been compared to a parametric IRT model performance. For an initial model comparison, all 23 binary variables have been used for performance comparison. The performance of all three IRT models is shown below, where

- Model 1: parametric IRT model, using R package ltm
- Model 2: nonparametric IRT model, using R package sirt
- Model 3: nonparametric IRT model with log-linear distribution smoothing for three moments respectively, using R package sirt

Table 6: Item response theory model comparison for parametric and non-parametric performance

	Item response theory models		
	Model 1	**Model 2**	**Model 3**
AIC	5,036,543	4,978,797	4,977,252
BIC	5,037,148	4,985,144	4,983,600
RSME	18.3407	0.8508	0.8143

Table 6 shows clearly that the nonparametric IRT models Model 2 and 3 outperform the parametric Model 1 due to smaller AIC and BIC and RSME in the latter

two models than Model 1. Applying a three-moment smoothing term does not change the distribution parameters of the IRT result but improves the model performance given further reduced RSME and information criteria (Xu & von Davier, 2008). Hence, we will move forward choosing Model 3, the nonparametric model with log-linear kernel distribution smoothing for three moments respectively.

3.2.2.2. Item Selection

As a result from the feature selection section 3.1.1. we know that the best number of relevant variables lays between 5 and 13. To explore which variables to include in the IRT model, we start with the 13 variables identified in section 3.1.1. with loosened cutoff criteria and reduce the number of variables step by step until the model does not further improve.

Table 7: Nonparametric IRT model comparison for different items from feature selection

Step number	Number of variables	AIC	BIC	RSME
0	23	4977252	4983600	0.8143
1	13	4870833	4874420	0.8153
2	12	3090255	3093567	0.8944
3	12	4026656	4029968	0.8378
4	12	4394784	4398096	0.8446
5	12	4761230	4764542	0.8176
6	12	4703253	4706565	0.8139
7	11	4666704	4669740	0.8138
8	11	4339160	4342196	0.8128
9	10	4302610	4305370	0.8128
10	9	4281952	4284436	0.8131
11	9	4274145	4276629	0.8144
12	9	4269695	4272178	0.8137
13	9	4185812	4188296	0.8141

The decision steps for the different models shown in Table 7 are defined below:

- Step 0: initial model, Model 3 from section 3.2.2.1.: implement nonparametric

IRT model starting with 13 binary complication variables from Table 4 in section 3.2.1.4. which summarizes all feature selection methods with loosened cutoff values: Maternal Transfusion, Perineal Laceration, Ruptured Uterus, Unplanned Hysterectomy, Admit to Intensive Care, Assisted Ventilation (immediately), Assisted Ventilation > 6 hrs, Admission to NICU, Surfactant, Antibiotics for Newborn, Cyanotic Congenital Heart Disease, Cleft Lip w/ or w/o Cleft Palate, Hypospadias

- Step 1: because of higher AIC and BIC than in the previous step, thus the model worsened, we remove *Admission to NICU*
- Step 2: Compared to Step 1, a higher RSME implies worse model, go back to Step 1 and remove *Assisted Ventilation (immediately)*
- Step 3: Compared to Step 1, a higher RSME implies worse model, go back to Step 1 and remove *Antibiotics*
- Step 4: Compared to Step 1, a higher RSME implies worse model, go back to Step 1 and remove *Surfactant*
- Step 5: Compared to Step 1, a higher RSME implies worse model, go back to Step 1 and remove *Assisted Ventilation > 6 hrs*
- Step 6: Compared to Step 1, a lower RSME implies a better model, move forward with this model and remove *Hypospadias*
- Step 7: Compared to Step 6, a higher RSME implies worse model, go back to Step 6 and remove *Perineal Laceration*
- Step 8: Move forward with this model, based on lower AIC, BIC and RMSE implying an improved scenario we go ahead and try model without *Hypospadias*

- Step 9: Judging from residuals compared with previous models move forward with this model, based on high item pair correlation try model without *Ruptured Uterus*
- Step 10: Compared to Step 9, a higher RSME implies worse model, go back to Step 9 and remove *Cyanotic Congenital Heart Disease*
- Step 11: Compared to Step 9, a higher RSME implies worse model, go back to Step 9 and remove *Cleft Lip w/ or w/o Cleft Palate*
- Step 12: Compared to Step 9, a higher RSME implies worse model, go back to Step 9 and try model without *Surfactant*
- Step 13: Compared to Step 9, a higher RSME implies worse model; every useful combination tried, no further model improvement possible. Go back to Step 9 and keep model 9

Based on the model comparison, we move forward with the results from model 9 which shows the lowest RMSE. Therefore, we have the following 10 items included in the nonparametric IRT with log-linear kernel distribution smoothing for three moments respectively: Maternal Transfusion, Ruptured Uterus, Unplanned Hysterectomy, Admit to Intensive Care, Admission to NICU, Assisted Ventilation (immediately), Surfactant, Antibiotics for Newborn, Cyanotic Congenital Heart Disease, and Cleft Lip with or without Cleft Palate. The distribution of the latent trait of the final model as described in

the previous section is shown in Figure 4:

mean	0.99
sd	0.62
skewness	2.84
kurtosis	7.28

Figure 4: Latent trait distribution for item response theory model

Figure 4 clearly shows that the continuous latent trait result from binary variables follows a skewed distribution. In fact, the distribution resembles a gamma distribution which creates a challenge for the normality approximation. The range of this latent trait lays between -0.1948 and 5.2576, whereas the majority of cases exhibits a score of -0.1948. Hereafter, we refer to this latent trait as IRT variable or IRT outcome and treat it as another numerical variable in the dataset.

3.3. Preprocessing Numerical Variables

In this section, we will give an overview on the distributions of the numerical variables, and prepare the variables for further transformation. We need to perform normalization transformation for factor analysis and SEM modeling, because their estimation methods require approximated normal distribution of the observed variables.

3.3.1. Descriptive Statistics

Table 8: Descriptive statistics of numerical variables for birth delivery data 2018

variable	mean	std	skewness	kurtosis
Birth weight (g)	3264.17	586.36	-0.8307	5.6298
Mother's age	29.01	5.8	0.0799	2.5246
Combined gestation	38.6	2.43	-1.8732	12.3447
Apgar score after 5 minutes	8.8	0.78	-5.3282	42.7834
Apgar score after 10 minutes without non measured	6.02	2.65	-0.8882	2.6726
Apgar score after 10 minutes	87.09	8.57	-9.3712	88.9213

Table 8 shows descriptive statistics of all selected numerical variables. According to ICD-10 (World Health Organization, 2004a), low birth weight is defined as less than 2,500 grams. The majority of the babies born in 2018 is above this cutoff. Babies born between the thirty-seventh and fortieth gestation week count as full-term birth. In our dataset we observe a mean gestation time of 38.6 weeks. The average birth weight for full-term babies is 3,200 grams (University of Rochester Medical Center Rochester, 2020). This general description matches our dataset where we observe slightly heavier average birth weight of 3264 grams. With an average Apgar 5 of 8.8, most babies are born healthily. Furthermore, the average Apgar 10 is 87, because Apgar 10 is only measured for babies with Apgar 5 of lower than 6. If an Apgar score did not get remeasured after 10 minutes (Apgar 10) then a value of 88 is noted (National Center for Health Statistics, 2019). To give a more precise picture about the score, both versions (including 88 and not including 88) are included in the data description in Table 1.

Furthermore, we calculated skewness and kurtosis for each of the continuous variables. The R package moments was used to calculate the kurtosis. Therefore, the

calculated kurtosis under the null hypothesis of normality would be 3 (Lukasz Komsta, 2015). In the dataset, only mother's age is close to be normally distributed, the other variables show deviation of normality. In particular, the Apgar scores are highly skewed which can be explained by the fact that most newborns achieve a high Apgar score in the dataset and therefore are generally born in a good condition.

To evaluate the variable distributions, we plot histograms with normal density graphs as shown below. Judging from the histograms and density plots, the variables *birth weight, mother's age*, and *combined gestation time* are close to symmetry. The only exception are the Apgar scores, where Apgar 5 is highly negative skewed and Apgar 10 shows an almost bimodal distribution, whereas the two peaks are around 0 (dead) and 6. For each variable, the histogram with density plot is shown in Figure 5 to 9:

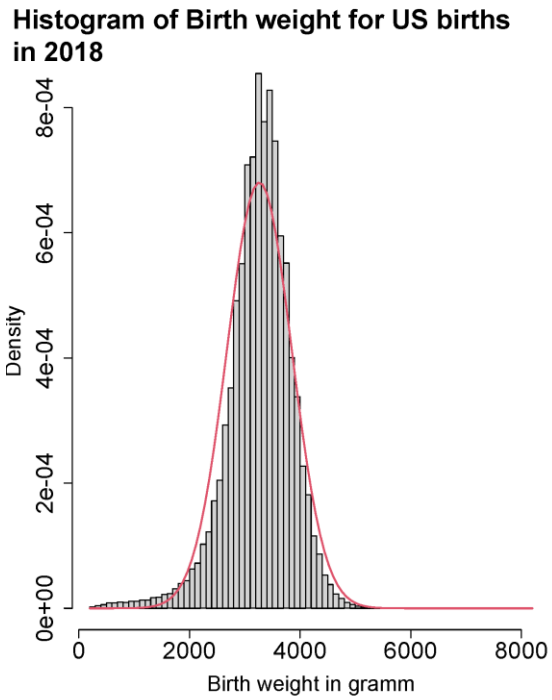

Figure 5: Histogram with density plot for birth

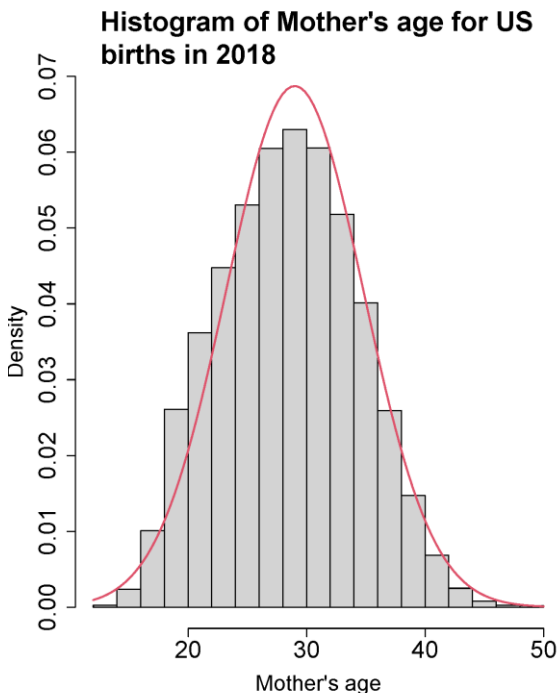

Figure 6: Histogram with density plot for mother's age 2018

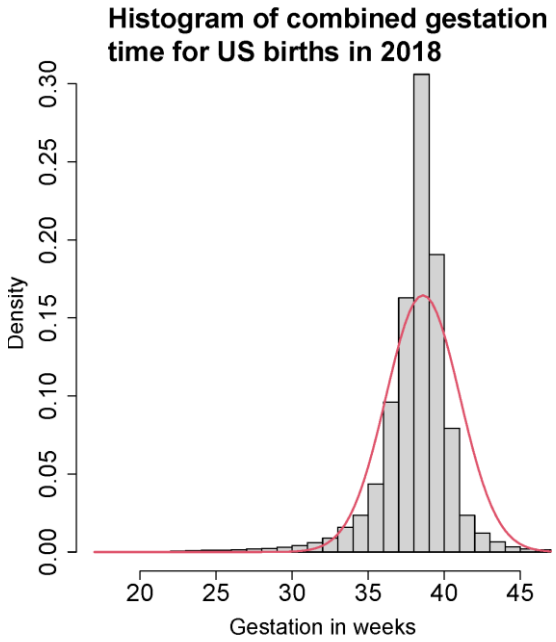

Figure 7: Histogram with density plot for

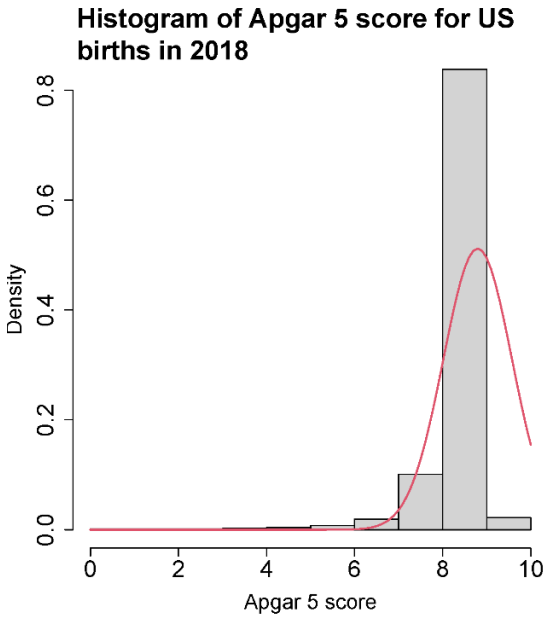

Figure 8: Histogram with density plot for

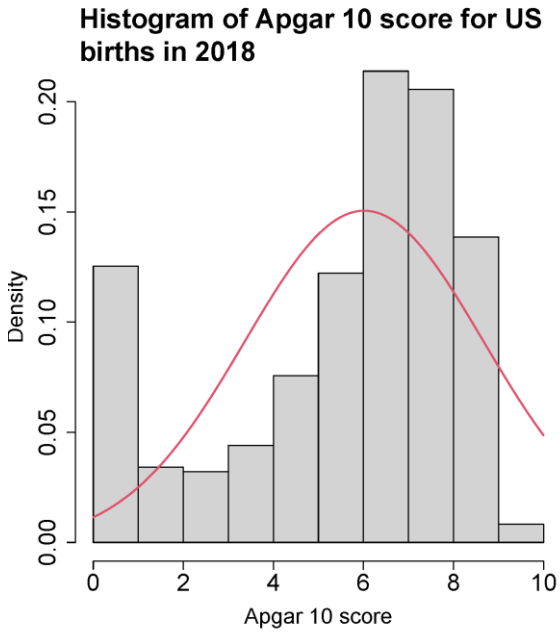

Figure 9: Histogram with density plot for Apgar

Figure 9 includes only measured Apgar 10 scores and excludes the cases in which a Apgar 10 score has not been necessary because of a sufficient Apgar 5 score.

3.3.2. Association between Numerical Observed Variables

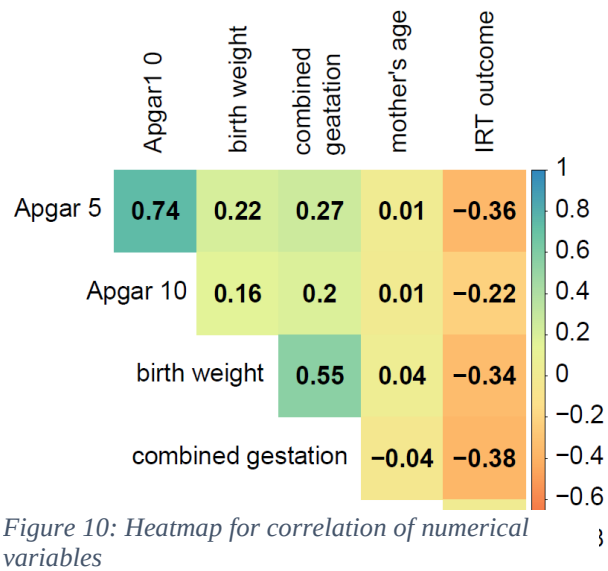

Figure 10: Heatmap for correlation of numerical variables

To investigate linear association between the observed variables, we present a heatmap in Figure 10. A strong positive correlation (0.74) between *Apgar score 5* and *Apgar score 10* is observed. This is due to the fact that the Apgar scores measure the baby from the same perspective; once after 5 minutes of delivery, again after 10 minutes if their Apgar 5 was lower than 7 (National Center for Health Statistics, 2019). Additionally, a moderate positive correlation (0.55) is observed between *birth weight* and *combined gestation time*. Furthermore, we detect a mild negative correlation between the IRT outcome and Apgar 5 (-0.36) which is due to the fact that Apgar 5 gives a score for the babies' ability to thrive while the IRT outcome summarizes the birth delivery complications. We also notice expected mild negative correlations between birth weight and IRT outcome (-0.34), and combined gestation time and IRT outcome (-0.38). Babies born prematurely often require assisted ventilation as their lungs may not be fully developed yet. Babies with low birth weight need more treatment in comparison to full-term birth deliveries; moreover, it is common procedure in nursing guidelines to administer babies born pre-term or with low birth weight into the NICU (Chakkarapani et al., 2020; Stanford Children's Health, 2021).

Table 8 describes the relationship between Apgar 5 and Apgar 10 in more detail.

It describes the 2-dimensional density and visualizes the transition from a baby's Apgar scores after 5 minutes (x-axis) to the Apgar score after 10 minutes (y-axis). The numbers highlighted in a bold font display the highest density in its category while the yellow cells describe a stagnation of the baby's potential of thriving. If the Apgar score 10 would not improve but stay as measured after 5 minutes, we would expect a diagonal cluster of density. However, this table shows that most babies improve their chance of survival (Apgar score) after 10 minutes. The second half of the table, Apgar scores 5 from values 6 to 10 describes the healthier babies and there is no need to remeasure their score after 10 minutes since their chances of survival are fairly high. When we take a closer look at the yellow cells we notice that the majority of babies with an Apgar score of 4 or 5 actually improve their score after 10 minutes to a value of 6 or even higher. On the other side, a majority of newborns with an Apgar 5 score of 0 to 2 fail to improve their Apgar score 10 minutes after being born and still struggle with survival.

Table 9: Relationship between initial Apgar score after 5 minutes and remeasured Apgar score after 10 minutes in percentage

Apgar score 10	Apgar score 5											
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	Total
0	52.49	31.16	7.96	2.84	1.66	3.88	0	0	0	0	0	100
1	3.76	82.11	10.03	2.44	0.87	0.79	0	0	0	0	0	100
2	6.69	21.97	59.37	8.17	1.97	1.83	0	0	0	0	0	100
3	5.24	26.18	25.58	36.65	3.89	2.47	0	0	0	0	0	100
4	2.72	15.31	24.69	26.54	27.52	3.22	0	0	0	0	0	100
5	1.02	11.9	15.77	24.53	23.07	23.71	0	0	0	0	0	100
6	0.33	7.03	12.14	18.76	28.76	32.97	0	0	0	0	0	100
7	0.21	4.98	8.43	14.21	24.54	47.63	0	0	0	0	0	100
8	0.23	4.33	8.51	12.24	23.35	51.34	0	0	0	0	0	100
9	1.13	5.2	9.22	11.16	22.62	50.68	0	0	0	0	0	100
10	2.62	4.07	4.36	7.85	11.05	70.06	0	0	0	0	0	100
88	0	0	0	0	0	0	**0.79**	**1.95**	**10.24**	**84.8**	**2.22**	100
Percent of total	0.03	0.17	0.14	0.16	0.22	0.38	0.78	1.93	10.12	83.86	2.19	**100**

As a last step before moving forward to the factor analysis and SEM modeling, we assess the association of the continuous variables. We assume a linear relationship between observed variables and latent variable. Therefore, we present a scatterplot matrix (Figure 11) to check if there is any nonlinear relationship between the observed variables which could endanger the linearity assumption.

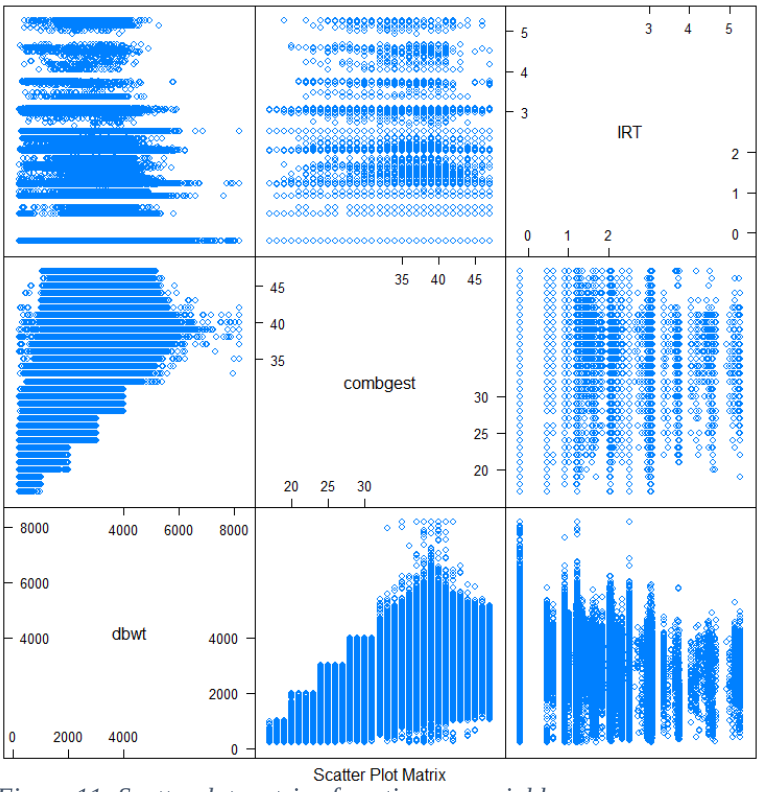

Figure 11: Scatterplot matrix of continuous variables

Based on the heatmap in Figure 10, we are already aware of the strong linear correlation between combined gestation time and birth weight. Besides this, there is a mild correlation between birth weight and the IRT result. In addition, we do not notice any nonlinear relationship and thus support our linearity assumption of the observed variables.

3.3.3. Rescaling Variables Apgar 10 and Birth Weight

If the Apgar 10 score was not performed due to a good Apgar score 5 (6 or higher), a value of 88 is coded for Apgar 10. In order to bring the variables Apgar 10 closer to the scale, we perform an explanatory factor analysis excluding birth weight, since it is correlated with combined gestation time. The Apgar 5 score was kept even though it has a correlation with Apgar 10, yet Apgar 10 still provides additional information to the previous measured Apgar 5 score. This leads to the following model:

Latent variable =~ Apgar 5 + Apgar 10 + combined gestation

As shown in the table below, the best value for the replacement of the missing

Apgar 10 score is the value of 12 according to its loading into the factor model and its uniqueness. A higher factor loading indicates a greater importance of the variable for predicting the latent variable.

Table 10: Decision steps for missing Apgar 10 score replacement using explanatory factor analysis model

# step	Decision steps	New Apgar 10 value	Explained variance	Eigen-values/ sum of squared loadings	Apgar 5 factor loading	Apgar 10 factor loading	Gestation time factor loading	Unique-ness of Apgar 10
0	Original data number	88	0.539	1.618	0.997	0.742	0.269	0.450
1	Starting point: Smallest possible number (Apgar score ranges from 0-10)	11	0.508	1.525	0.894	0.797	0.300	0.364
2	Check higher apgar10 value: increase value in step 1 by 4	15	0.530	1.591	0.937	0.794	0.287	0.369
3	Check higher apgar10 value: increase value in step 2 by 8	23	0.537	1.611	0.967	0.774	0.278	0.401
4	Check higher apgar10 value: increase value in step 3 by 16	39	0.539	1.616	0.985	0.756	0.272	0.429
5	Uniqueness almost as high as in original model, check values between step 1 and 2	13	0.524	1.571	0.920	0.800	0.292	0.361
6	Uniqueness lower than step 1 and 2,	12	0.518	1.554	0.909	0.800	0.295	0.360

# step	Decision steps	New Apgar 10 value	Explained variance	Eigen-values/ sum of squared loadings	Apgar 5 factor loading	Apgar 10 factor loading	Gestation time factor loading	Unique-ness of Apgar 10
	decrease the values in step 5 by 1							
7	Uniqueness in step 6 lowest so far, increase the value in step 6 by 2	14	0.528	1.583	0.929	0.797	0.289	0.364
Result: replacing Apgar10 value 88 with 12 gives best results regarding uniqueness of apgar10 variable								

Based on the models above and judging from the lowest uniqueness and the highest factor loading, we will use 12 instead of 88 for not conducted Apgar scores after 10 minutes.

Because the birth weight variable is in the unit of grams, it has a larger scale than the other variables in the model. Therefore, we divide the birth weight data by 453.592 to measure it in the unit of pound and therefore bring this variable closer to the scale of the others.

3.3.4. Normalization of Numerical Variables

Since multiple statistical methods require a normality assumption for variables, we need to transform skewed variables in order to approach a more normal distribution. Different normalization methods are described based on the skewness of the variable to normalize skewed distribution. The most common ones are the log transformation which is recommended if the data is moderately to severe skewed. For less skewed data, the square-root is recommended while severe skewed data may be approach normality better

with an inverse transformation. Differences in the transformation are based on the direction of skewness, negatively or positively skewed (Kassambara, 2018).

Since R is a common statistical software various packages for normality transformations are available. A very powerful method is the newly method called "Gaussianize" which is available in the LambertW package in R (Georg, 2020). It utilizes an inverse transformation which allows to correct heavy tails. Furthermore, it utilized the Tukey distribution for developing a parametric function which can handle skewed and heavy-tailed data jointly (Goerg, 2015). Table 10 shows the power of this function compared to other common normalization methods such as boxcox, simple inverse, square-root, and log transformation. The function *transformTukey* from R package *rcompanion* (Mangiafico, 2021) was not able to handle datasets above a sample size of 5000. The boxcox transformation was performed using the R package geoR (Jr et al., 2020). Another package by R for normality transformation is the package *bestNormalize* (Peterson, 2017), which uses multiple transformation methods and returns the best fit. It chooses from ordered quantile normalization, which consists of a rank-mapping plus shifted logit approximation, boxcox and Yep-Johnson transformation, and in total three types of Lambert WxF transformations. All these transformations are automatically compared by the function in the background to the original data skewness and kurtosis. Finally, the best transformation approach with the resulting converted data is returned which is closest to normality (Peterson, 2017).

We test different transformation methods and their combinations to find a close approximation of each variable to normality. One exception is Apgar 10. It is difficult to transform it, since the majority of cases exhibit a not-performed value. Even with the data

transformation of this value from 88 to 12, the variable could not be normalized. The closest approximation was log (13– Apgar 10) where the results are still far from normality. However, the replacement of Apgar 10 score 88 provided a wider range of values than using the log function. Therefore, we did not normalize Apgar 10 but replace its missing values with the value 12. After applying the Gaussianize function, we did not move forward with normalization methods for birth weight and combined gestation because we observe a satisfactory normal approximation with a skewness of almost zero and a normal kurtosis.

Table 11: Result of applied normalization methods

	Variable							
	Birth weight		**Combined gestation**		**Apgar 5**		**Apgar10 *****	
Method	skewness	kurtosis	skewness	kurtosis	skewness	kurtosis	skewness	kurtosis
Original	-0.831	5.63	-1.873	12.345	-5.328	42.783	-11.991	162.09
sqrt(max(x+1) – x)	0.487	5.287	0.564	8.771	3.58	23.62	10.623	122.244
log10(max(x+1) – x)	-1.048	89.232	-1.231	14.592	1.956	13.751	9.873	101.81
1/(max(x+1) – x)	580.554	338487.7	10.824	184.412	1.488	14.593	-9.416	90.042
boxcox	0.195	4.423	0.293	6.905	-1.414	8.905	-9.5	92.036
Gaussianize()*	0	3	0.003	3	-0.002	13.813		
Transform Tukey()	sample size too large (n=5000 max)							
bestNormalize comparison**	boxcox		boxcox		Standardized exp(x)		Standardized	Log_b(x + a)
bestNormalize calculation					1.645	14.073	-39.267	1767.116
log(constant – x)					-6.37	42.342	9.391	89.414
first boxcox, then Gaussianize()					-1.257	5.317		
first boxcox, then Gaussianize() twice					-1.258	5.316		
First Gaussianize() then boxcox					0.247	13.924		
Gaussianize skewed + symmetric-heavy tails					Error	Error		
exp(x), then Gaussianize()					Error	Error		
exp(x), then boxcox()					-0.215	13.807		

*on birth weight transformed into lbs, Apgar scores +1 to avoid zeros

** comparing arcsinh(x), Box-Cox, Exp(x), Log_b(x+a), no transform, sqrt(x + a), Yeo-Johnson

*** 88 replaced with 12

3.3.5. Artificial Data for IRT Model Normalization

The normalization methods as described in section 3.3.3. failed for the latent trait variable of the item response model. None of the methods could convert it to normality. This is caused by the highly asymmetric data which form an approximate gamma distribution. As an approach to reduce the skewness of the variable an artificial data approach has been chosen: First, we filtered the IRT variable for values which are bigger than 0 since these are causing the heavy right tail. There are in total 427,984 cases (11.35%) of a score above 0. These scores have been multiplied by -1, to create a left tail, thus mitigating the skewness of the data. Values for the other numerical values have been simulated based on their variable distributions in the original dataset. This simulation step adds 427,984 simulated birth delivery cases which has been flagged as non-original data to identify them from the real cases. The histogram of the data in Figure 12 shows the symmetric IRT variable after adding the artificial data. Originally the dataset has 3,769,386 observations, by adding the artificial data to remove the skewness, we have a resulting in a total dataset of 4,197,370 observations. We keep the simulated cases until the latent variable has been estimated. After the latent variable estimation, the simulated cases will be removed and only the original cases will be kept for further analysis.

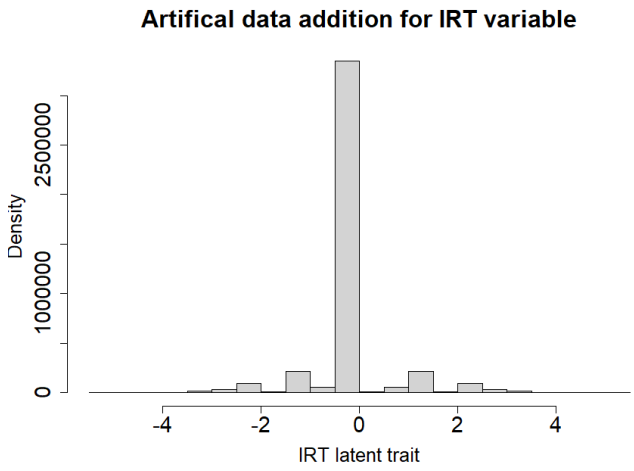

Figure 12: Histogram of IRT variable with artificial data

4: RESULTS

In this section, we present the estimation results of the latent variable, the birth quality, from several methods, namely factor analysis, SEM and MCMC. We also consider possible multiple modals of the latent variable using clustering and the selected SEM estimation. For all our computation, we use an Intel® Xeon® Gold 5220 CPU with 2.19 GHz (2 processors), 18 cores and 128 GB RAM.

4.1. Factor Analysis Results

Explanatory factor analysis is conducted in R using the psych package (Revelle, 2017). To select the best model, we create an initial model with all numerical modify the model accordingly to the model-fit:

Table 12: Explanatory factor analysis model comparison

	model 1	model 2	model 3	model 4
factor loadings				
Apgar 5	0.5	0.5	NA	NA
Apgar 10	0.42	0.42	0.22	0.29
combined gestation time	0.5	0.5	0.66	NA
birth weight	0.49	0.49	0.65	0.41
continuous variable from IRT	-0.57	-0.57	-0.45	-0.67
mother's age	0	NA	NA	NA
communality (h^2)				
Apgar 5	0.25	0.25	NA	NA
Apgar 10	0.18	0.18	0.05	0.085
combined gestation time	0.25	0.25	0.43	NA
birth weight	0.24	0.24	0.42	0.171
continuous variable from IRT	0.32	0.32	0.20	0.451
mother's age	0.0000052	NA	NA	NA
Goodness-of-fit indices				
TLI	0.444	0.354	0.874	-Inf
RMSEA	0.158	0.208	0.088	NA
BIC	5111160	4064274	236041	NA

The decision steps and their resulting four models are described below. To start with, we include all numeric variables in Model 1: mother's age, Apgar 5, Apgar 10, birth weight, combined gestation time and IRT outcome

Step 1: Based on the low influence of mother's age as shown in no factor loading and its low communality, we remove mother's age

Step 2: Compared to Step 1, a lower BIC implies better model, but lower TLI and higher RMSEA imply the model worsened; TLI and RMSEA are sensitive for correlation, so we remove Apgar 5 in Model 2, because of its linear correlation with Apgar 10 and IRT outcome.

Step 3: Compared to Step 2, TLI increased and RMSEA and BIC lowered which implies better model fit, so we remove combined gestation based in Model 3, based on their linear correlation with birth weight and IRT outcome

Step 4: Compared to Step 3. Not enough input parameters for model fit evaluation , (model needs more than 3 indicators otherwise TLI = -Inf), therefore Model 4 is not identified (TLI = -Inf), go back to Step 3 and keep Model 3

Resulting from the decision steps and the model comparison table above, the best explanatory factor analysis model is model 3 with the following indicators to quantify the birth delivery quality: Apgar 10, birth weight, combined gestation time, IRT outcome. The distribution of the newly estimated variable to quantify birth delivery is shown in Figure 13.

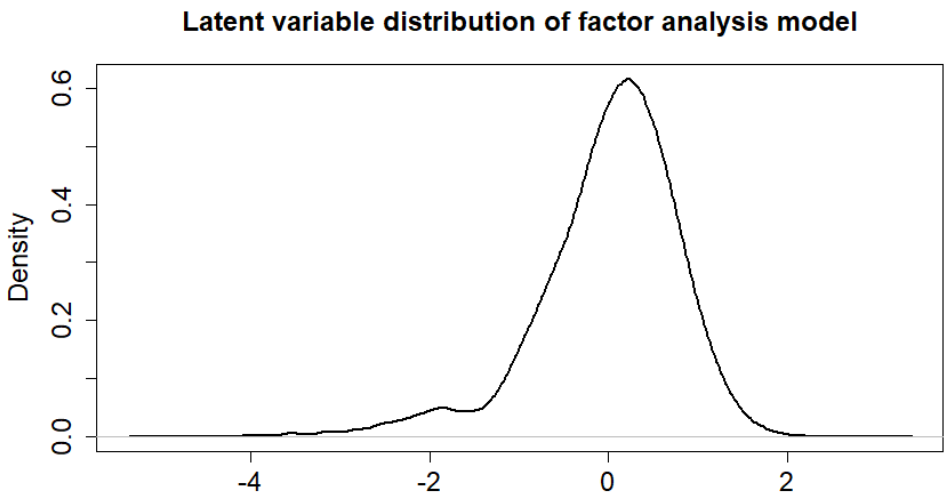

Figure 13: Density plot of latent variable estimated by factor analysis model

We notice a negatively skewed distribution for the quality of birth delivery (the latent variable), with a median of 0.1 and a mean of 0 (sd 0.8). The latent variable ranges from -5.28 to 3.29 and exhibits a skewness of -1.05 and a kurtosis of 2.19. Since most birth deliveries do not exhibit much negative effect, we can conclude that a high score range shows good quality of birth deliveries whereas the negative scores represent the birth deliveries with complications.

Since we include artificial data for the result, we want to compare the model performance between the original dataset and the dataset with artificial data from Section 3.3.4. The result is presented in Table12 and uses the final model from the explanatory factor analysis result above.

We notice stronger factor loadings in the original dataset and a higher communality. Considering the residuals, BIC, and TLI, we notice that the artificial data clearly improved the model fit, which is explained by the normality assumption of factor analysis models (Jöreskog, 1977; Wall et al., 2012). We also notice that the latent variable estimation for the original data is more skewed than the latent variable

estimation resulted from the dataset with artificial data.

Table 13: Comparison of latent variable estimation on original and artificial data

	Original data	artificial data
factor loadings		
apgar10	0.22	0.19
combined gestation time	0.66	0.62
birth weight	0.65	0.62
IRT outcome	-0.45	-0.33
communality (h^2)		
apgar10	0.05	0.037
combined gestation time	0.43	0.387
birth weight	0.42	0.383
IRT outcome	0.20	0.109
Goodness-of-fit indices		
Tucker Lewis Index (TLI)	0.874	0.913
RMSEA	0.088	0.06
BIC	236041	122807.9
Skewness parameters for the latent variable estimation		
skew	-0.98	-0.83
kurtosis	1.88	1.68

We move forward with using the artificial data for the estimation methods with underlying normality assumptions for model parameters. After the latent variable is estimated, the simulated data is removed from the dataset and we provide summary statistics of the latent variable based on the original dataset.

4.2. Structural Equation Model Results

Similar as the factor analysis model, SEM assume a linear association between observed and latent variables but extends factor analysis by allowing to specify residual covariance. We conduct SEM modeling by using the lavaan package in R (Rosseel, 2012), which assumes a correlation between observed and latent variables. The default estimator for the lavaan package is the maximum likelihood estimation which relies on a

normality assumption for the observed variables (Rosseel, 2012). SEM models can handle a mild level of nonnormality, that is, variables with a skewness between -3 and 3 are acceptable, and a kurtosis between -10 and 10 is tolerated. For variables outside the tolerance the model still works but loses reliability (Griffin & Steinbrecher, 2013).

To account for skewness in the dataset, lavaan offers a maximum likelihood estimation with robust standard errors plus Satorra-Bentler scaled test statistic (further referred to as MLM) and a weighted least squares estimation (WLS) which is also called asymptotically distribution-free estimation method (Rosseel, 2012). Since our dataset is not exactly normally distributed after the normalizing transformation, we utilize both estimation methods to account for the remaining level of nonnormality in the dataset.

4.2.1. SEM Model with Maximum Likelihood Estimation and Robust Standard Errors plus Satorra-Bentler Scaled Test Statistic (MLM)

For our first SEM model, we use SEM MLM estimation for our latent variable birth delivery quality. The model is based on the normalized dataset resulting from section 3.3.3. We present the performance measures in Table 14.

Table 14: SEM model with maximum likelihood estimation, robust standard errors, and Satorra-Bentler scaled test statistic (MLM)

	1	2	3	4	5	6	7	8
Chi square	341539.643	101837	29512.9	11789.3	5276.085	3946.254	386.629	21.725
degreeof freedom	9	8	7	6	5	3	2	1
p value	0	0	0	0	0	0	0	0
Robust CFI	0.645	0.924	0.981	0.992	0.996	0.997	1	1
Robust TLI	0.409	0.858	0.959	0.98	0.989	0.989	1	1
AIC	77839549	77283404	77171056	77148695	77140232	50492165	50486010	50485652
BIC	77839708	77283577	77171242	77148894	77140444	50492324	50486182	50485838
Robust RMSEA	0.137	0.067	0.036	0.025	0.019	0.023	0.007	0.002
RMSEA lower bound CI*	0.136	0.067	0.036	0.025	0.018	0.022	0.006	0.001
RMSEA upper bound CI*	0.137	0.067	0.036	0.026	0.019	0.023	0.007	0.003
Robust SRMR	0.074	0.039	0.017	0.012	0.008	0.008	0.002	0

* 90% confidence interval (CI)

The latent variable is named Health in the SEM models. We start with an initial model using all numerical indicators for Model 1:

Health =~ combined gestation + birth weight + Apgar 5 + IRT +Apgar 10 + mother's age

Residual (co)variances:

None

Step 1: Model 1 indicates us to implement a residual covariance path between Apgar 5 and Apgar 10 based on highest modification index. We do so in Model 2:

Health =~ combined gestation + birth weight + Apgar 5 + IRT +Apgar 10 + mother's age

Residual (co)variances:

Apgar 5 ~~ Apgar 10

Step 2: We compare Model 1 and 2, the latter has lower residuals and larger goodness-of-fit indices, so Model 2 is better than Model 1. Based on the highest modification index for the residual covariance path between birth weight and combined gestation, we implement this path in

Model 3:

Health =~ combined gestation + birth weight + Apgar 5 + IRT +Apgar 10 + mother's age

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

Step 3: We compare Model 2 and 3, the latter has lower residuals and larger goodness-of-fit indices, so Model 3 is better than Model 2. Based on the highest modification index for the residual covariance path between mother's age and birth weight, we implement this path in Model 4:

Health =~ combined gestation + birth weight + Apgar 5 + IRT +Apgar 10 + mother's age

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

mother's age ~~ birth weight

Step 4: We compare Model 3 and 4, the latter has lower residuals and larger goodness-of-fit indices, so Model 4 is better than Model 3. Based on the highest modification index for the residual covariance path between mother's age and combined gestation, we implement this path in Model 5:

Health =~ combined gestation + birth weight + Apgar 5 + IRT +Apgar 10 + mother's age

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

mother's age ~~ birth weight

combined gestation ~~ mother's age

Step 5: We compare Model 4 and 5 and notice that even with residual covariance paths for mother's age, the variance of mother's age does not change. It is very high with a value of 33, therefore we remove mother's age and its paths resulting in Model 6:

Health =~ combined gestation + birth weight + Apgar 5 + IRT +Apgar 10

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

Step 6: We compare Model 5 and 6, the latter has lower residuals and larger goodness-of-fit indices, so Model 6 is better than Model 5. Based on the highest modification index

for the residual covariance path between Apgar 5 and IRT outcome, we implement this path in Model 7:

Health =~ combined gestation + birth weight + Apgar 5 + IRT +Apgar 10

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

Apgar 5 ~~ IRT

Step 7: We compare Model 6 and 7, the latter has lower residuals and larger goodness-of-fit indices, so Model 7 is better than Model 6. Based on the highest modification index for the residual covariance path between birth weight and Apgar 10 outcome, we implement this path in Model 8:

Health =~ combined gestation + birth weight + Apgar 5 + IRT +Apgar 10

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

Apgar 5 ~~ IRT

Birth weight ~~ Apgar 10

Step 8: We compare Model 7 and 8, the latter has lower residuals and larger goodness-of-fit indices, so Model 8 is better than Model 7. Our model has with one remaining degree of freedom no more degrees of freedom left for implementing more residual covariance paths. We keep this model.

The density plot for the SEM model 8 which has been proven as best MLM estimation model is shown in Figure 14. We clearly see the majority of cases between 0 and 0.5, while we see a second but much smaller peak around -1.5 which would be severely impacted cases of birth delivery. From the skewness and kurtosis parameters in for the SEM MLM estimation in Table 14, we see a negatively skewed distribution. About 50% of all cases do have a score above zero.

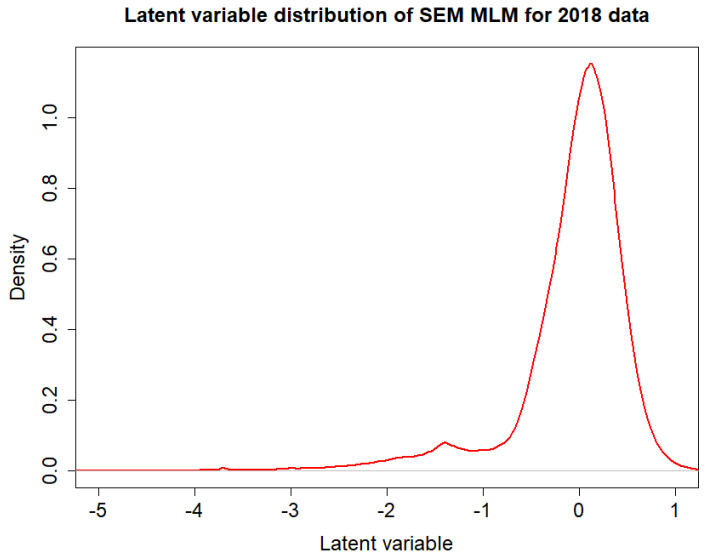

Figure 14: Density plot of latent variable estimated by SEM MLM

Table 15: SEM MLM latent variable distribution

Latent variable distribution	
mean	-0.08
median	0.05
sd	0.62
skew	-2.39
kurtosis	8.75

4.2.2. SEM Model with Weighted Least Squares Estimation (WLS)

Similar to the MLM SEM estimation, we create a SEM model using WLS estimation. For a more conservative estimation, we first use the model on the

unnormalized data and will then compare the outcome from the best SEM WLS model between unnormalized and normalized data. The results are presented in Table 16.

Table 16: SEM model with weighted least squares estimation (WLS)

	Model number							
	1**	**2****	**3**	**4**	**5**	**6**	**7**	**8**
Chi square	NA	NA	35634.7	345739	6507.877	4773.172	2560.822	426.605
degrees of freedom	NA	NA	7	6	5	3	2	1
p value	NA	NA	0	0	0	0	0	0
Robust CFI	NA	NA	0.897	0.966	0.981	0.985	0.992	0.999
Robust TLI	NA	NA	0.779	0.914	0.944	0.948	0.959	0.986
AIC	NA	NA	-	-	-	-	-	-
BIC	NA	NA	-	-	-	-	-	-
Robust RMSEA	NA	NA	0.037	0.023	0.019	0.021	0.018	0.011
RMSEA lower bound confidence interval*	NA	NA	0.036	0.022	0.018	0.02	0.018	0.010
RMSEA upper bound confidence interval*	NA	NA	0.037	0.023	0.019	0.021	0.019	0.011
Robust SRMR	NA	NA	0.088	0.088	0.087	0.099	0.042	0.008

*90% confidence interval

**The optimizer did not find a solution for model 1 and model 2

Model 1:

Health =~ combined gestation + birth weight + Apgar 5 + IRT + Apgar 10 + mother's age

Residual (co)variances:

None

Step 1: no solution has been found, implement residual covariance between Apgar 5 and Apgar 10 based on previous SEM MLM model in Model 2:

Health =~ combined gestation + birth weight + Apgar 5 + IRT + Apgar 10 + mother's age

Residual (co)variances:

Apgar 5 ~~ Apgar 10

Step 2: no solution has been found, implement residual covariance between combined gestation and birth weight based on previous SEM MLM model in Model 3:

Health =~ combined gestation + birth weight + Apgar 5 + IRT + Apgar 10 + mother's age

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

Step 3: The optimizer found a solution. Model 1 indicates us to implement a residual covariance path between mother's age and birth weight based on highest modification index. We do so in Model 4:

Health =~ combined gestation + birth weight + Apgar 5 + IRT + Apgar 10 + mother's age

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

mother's age ~~ birth weight

Step 4: We compare Model 3 and 4, the latter has lower residuals and larger goodness-of-fit indices, so Model 4 is better than Model 3. Based on the highest modification index for the residual covariance path between mother's age and combined gestation, we

implement this path in Model 5:

Health =~ combined gestation + birth weight + Apgar 5 + IRT + Apgar 10 + mother's age

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

mother's age ~~ birth weight

combined gestation ~~ mother's age

Step 5: We compare Model 4 and 5 and notice that even with residual covariance paths for mother's age, the variance of mother's age does not change. It is very high with a value of 33, therefore we remove mother's age and its paths resulting in Model 6:

Health =~ combined gestation + birth weight + Apgar 5 + IRT + Apgar 10

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

Step 6: We compare Model 5 and 6, the latter has lower residuals and larger goodness-of-fit indices, so Model 6 is better than Model 5. Based on the highest modification index for the residual covariance path between Apgar 5 and IRT outcome, we implement this path in Model 7:

Health =~ combined gestation + birth weight + Apgar 5 + IRT +Apgar 10

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

Apgar 5 ~~ IRT

Step 7: We compare Model 7 and 6, the latter has lower residuals and larger goodness-of-fit indices, so Model 7 is better than Model 6. Based on the highest modification index for the residual covariance path between birth weight and Apgar 10, we implement this path in Model 8:

Health =~ combined gestation + birth weight + Apgar 5 + IRT +Apgar 10

Residual (co)variances:

Apgar 5 ~~ Apgar 10

combined gestation ~~ birth weight

IRT ~~ birth weight

Birth weight ~~ Apgar 10

Step 8: We compare Model 7 and 8, the latter has lower residuals and larger goodness-of-fit indices, so Model 8 is better than Model 7. Our model has with one remaining degree of freedom no more degrees of freedom left for implementing more residual covariance paths. We keep this model. Figure 15 provides a density plot of the latent variable estimated by the SEM WLS method. We notice one smaller modal on the left side. Compared to the MLM estimation method the distribution seems smoother where we noticed two modals on the left hand side of Figure 15.

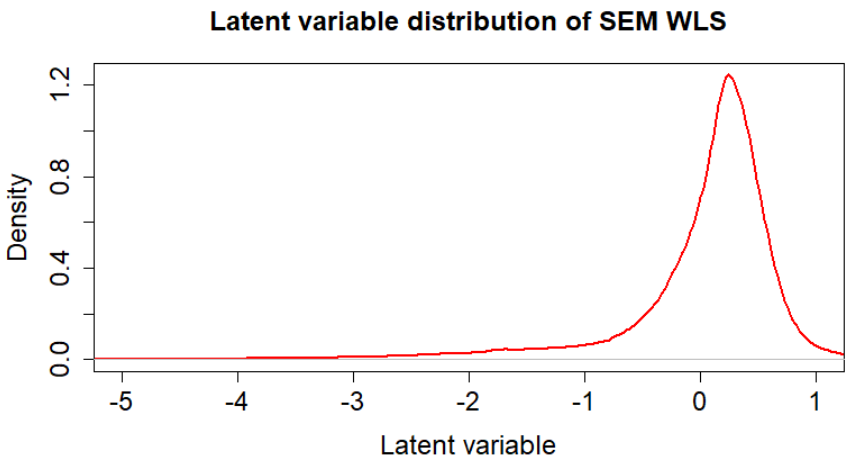

Figure 15: Density plot for latent variable estimated by SEM WLS

Similar to the SEM MLM estimation, we notice that the majority of cases exhibit a quality score above zero. In Table 16, the SEM WLS skewness and kurtosis parameters show an even higher values than the SEM MLM estimation.

Table 17: SEM WLS latent variable distribution

Latent variable distribution	
mean	0
median	0.29
sd	1.23
skew	-3.09
kurtosis	14.64

4.2.3. SEM Model Comparison

The best model out of the described SEM WLS estimated models is Model 8. In the MLM estimation we chose Model 8 as well. Since we used the raw data for evaluating the WLS model performance, we will apply the model to the normalized data to have a direct comparison to the SEM MLM model.

Table 18: SEM WLS and MLM model comparison

	SEM WLS model on original data	SEM WLS model on dataset with artificial data	SEM MLM model on dataset with artificial data
Chi square	426.605	526.044	21.725
degrees of freedom	1	1	1
p value	0	0	0
Robust CFI	0.999	0.999	1
Robust TLI	0.986	0.991	1
Robust RMSEA	0.011	0.012	0.002
RMSEA lower bound confidence interval*	0.010	0.011	0.001
RMSEA upper bound confidence interval*	0.011	0.013	0.003
Robust SRMR	0.008	0.003	0

* 90% CI

By comparing the best SEM WLS and MLM estimation models in Table 17, we clearly notice that the SEM MLM model outperforms the WLS estimator. Lower residuals and higher goodness-of-fit indices clearly demonstrate better results from the MLM estimation method which uses the transformed data to meet the normality assumptions. Even when we apply the SEM WLS model on the same data, it fails to improve the model fit.

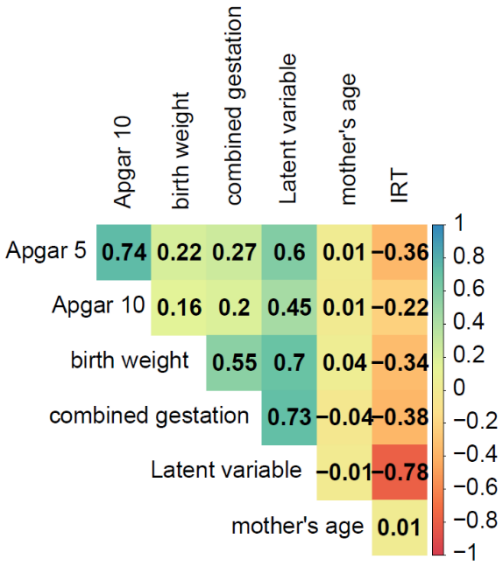

Figure 16: Heatmap of SEM MLM latent variable with observed variables for 2018

We provide a heatmap for the SEM MLM model in Figure 16. We can clearly see that the major indicator for our latent variable is our IRT outcome, which demonstrates a high negative correlation with our latent variable. Interesting to note is that mother's age is not correlated with any of our model variables.

4.3. MCMC

As described in the literature review, we found no literature which applied an MCMC model on a huge dataset as ours. Since MCMC is a parametric approach, we expect the calculation time to be slow. Other authors such as Fahrmeir and Raach use subsamples to speed up the estimation: they work with 170,000 survey samples but apply their MCMC model on only a subsample of 6,804 cases (Fahrmeir & Raach, 2007). To estimate our latent variable, we chose the brms package in R which works with the HMC algorithm and uses Stan as underlying platform. HMC algorithms can be seen as alternative algorithm to MCMC and have been become more popular in the more recent years because of their faster computation times (Monnahan et al., 2017).

First, we start with the model and prior specification. The brms package works with R as an interface, the user does specify the model in R while the brms package (Bürkner, 2017) uses C to calculate the model on the Stan platform. To estimate a latent variable in brms, we add an empty column to the dataset and declare this column as missing values for the model. We specify the latent variable model as follows:

```
fact_mod = bf(apgar5 ~ mi(Health)) +
bf(apgar10 ~ mi(Health)) +
bf(dbwt ~ mi(Health)) +
bf(combgest ~ mi(Health)) +
bf(IRT ~ mi(Health)) +
bf(Health | mi() ~ 0) + set_rescor(88escore = FALSE),
```

where Health is the latent variable, mi declares missing values, '~' stands for 'corresponds to', and bf stands for bayes formula. The general way to formulate the bayes

formula is "*y | subset(sub) ~ predictors*". Since our latent variable is declared as missing, the author of the brms package defines the SEM formulation with latent variable as stated above. Residual correlation in brms is currently only available for gaussian distributions. We initialize the sampling algorithm with an initial starting value of zero, three chains, 20,000 iterations and a small subsample of 1,000 randomly sampled cases. With only 1,000 cases, the computation time varies around 5,289.53 – 6,125.93 sec. Naturally, when setting the sample size higher, the computation time increases. We observe that with 2,000 samples, the computation time increases to 31,608.8 – 65,859.4 seconds. Moreover, a model with 5,000 samples does run for two – three days. This means it is impossible to use HMC as latent variable estimation for the whole dataset within a reasonable computation time.

However, using a fair estimated variable from SEM MLM estimation in section 4.2.1., as the initial prior distribution, we can speed up estimating the underlying distribution of the SEM MLM estimation with HMC. With 2000 iterations applied on a sample size of 500 for initial comparison of prior specified distributions for the hidden variable. In other words, here we assume the skew-normal distributed latent variable.

Table 19: Comparison of different distribution assumptions for latent variable estimated by MCMC

Hidden variable distribution assumption	WAIC estimate	SE
Skew normal	14532.2	208
Gaussian	2.93706e+11	1.03087e+11
Asymmetric Laplace	18769.3	124.5
Exgaussian	20624.1	121.9
Student t	19254.9	570

Since we include Apgar 5 in the HMC model, Apgar 10 has the lowest marginal information because only a few severe impacted babies get a remeasured Apgar score after 10 minutes. We analyze if Apgar 10 actually benefits our model or not by a model comparison using a sample size of 1000:

Table 20: Comparison between MCMC model with and without Apgar 10 variable

Model	LOOIC	SE (LOOIC)	WAIC	SE (WAIC)
Model with Apgar 10	52757212.3	7486511.9	5.29259e+11	1.1329e+11
Model without Apgar 10	17106.3	322.7	17074.6	320.3

Judging from the comparison table with and without Apgar 10 in Table 19, we clearly notice that LOOIC and WAIC are much lower in the model without Apgar 10, indicating a better model-fit. To verify this finding, we take a closer look at the pareto k estimates for each model:

Table 21: Comparison of Pareto k diagnostic values for MCMC models

		Model with Apgar 10		Model without Apgar 10	
Pareto k value		Count	Percent of total	Count	Percent of total
good	(-Inf, 0.5]	198	19.8%	970	97%
ok	(0.5, 0.7]	15	1.5%	9	0.90%
bad	(0.7, 1]	35	3.5%	17	1.70%
very bad	(1, Inf)	752	75.2%	4	0.40%

The pareto k values in Table 20 are computed from a 40,000 by 1,000 log-likelihood matrix. By analyzing the pareto k estimates in the model with Apgar 10, we notice 75.2% of all values are a very bad fit. In the model without Apgar 10 we reduce the very bad fitting pareto k estimates to 0.4% out of the total data. This supports our findings from the LOOIC and WAIC comparison (Table 18) between those models. Therefore, Apgar 10 is excluded from our HMC model and we adjust the remaining bad

fitting pareto k estimates with a loo moments match function in the brms package.

Since our SEM results showed at least two peaks in the distribution we take a closer look at the posterior distribution. For brms, we have to specify a distribution for the underlying population prior to the sampling algorithm, otherwise the computation may not converge. In Table 21, we try a mixture of two different distribution types and compare those with Model 4, our skew-normal specified posterior. The skew-normal assumption for the latent variable distribution has been further discussed above and is described in Table 18. In Table 21, we present the bimodal mixture distributions in comparison to Model 4:

Table 22: HMC comparison of bimodal distributions vs skew normal for posterior distribution

	Model 1	**Model 2**	**Model 3**	**Model 4**
Family 1	student-t	gaussian	gaussian	skew-normal
Family 2	skew-normal	skew-normal	student-t	-
LOOIC	1985 (SE 85.5)	1990.6 (SE 85.8)	1990.6 (SE 85.8)	2396 (SE 101.5)

We set a seed of 5 to have a direct model comparison, and use 1500 samples with 2500 iterations. Judging from the LOOIC, where a low LOOIC indicates a better model, we see that a bimodal distribution has a better model-fit than a simple skew-normal assumption. Furthermore, we observe that using a mixture of the distributions student-t and skew-normal outperforms the skew-normal assumption in model 4. Computation time of this model lays between 2-3 days for only 1500 samples.

Since the computation for a bimodal distribution is so complex, Rstudio randomly crashes in the middle of calculations. This limits our study to only a small subsample which is very unlikely representable of our dataset since we have only very few severe complications. However, we can conclude that the latent variable distribution does

exhibit a bimodal distribution and is not just skewed. We apply kmeans clustering to see if we can categorize the birth delivery quality into clusters for different quality.

4.4. Clustering

To analyze a possible multiple modal shape of the latent variable in more detail, we perform clustering before SEM to estimate the latent variable. We expect to show a significant difference between those clusters representing the modals in the latent variable scores found in the previous estimations. First, we create a scree plot to determine the optimal number of clusters:

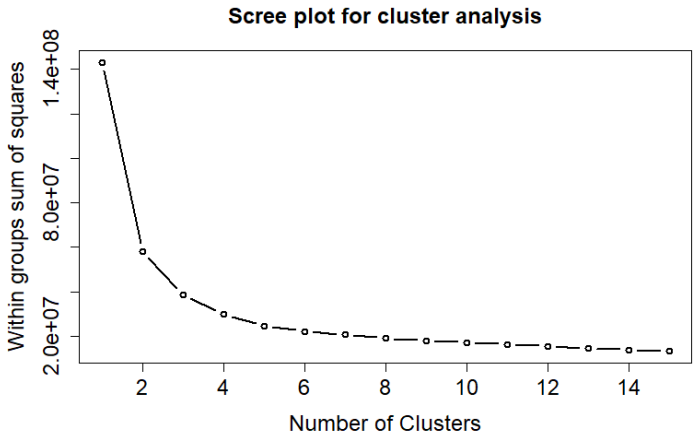

Figure 17: Scree plot for cluster analysis

Figure 17 shows the decreasing variance as the number of cluster increases. It is calculated based on the within groups sum of squares (WSS) estimation. We observe the steepest WSS reduction in 2 clusters. Judging from the scree plot, the optimal amount of clusters is either 2 or 3. We will try both options and compare their performance (Frushicheva, 2016; Woods & Edwards, 2011).

For the latent variable estimation of each cluster, we chose to apply both SEM estimators, MLM and WLS. First, we apply the clustering algorithm on the original data

without any transformation. Then, we split the dataset by cluster and apply the SEM model to each cluster of the dataset.

To test if the clusters differ from each other we apply the Kruskal-Wallis test, a nonparametric test by rank (Glen, 2021b). It is an alternative to the parametric one-way ANOVA, which we cannot apply since our latent variable is skewed in every cluster judging from our previous density plots. The Kruskal-Wallis test by rank extends the two-sample Wilcoxon test in a situation with more than two groups. The result (Table 22) shows that all clusters differ significantly from each other with a significance level of 0.05.

Table 23: Kruskal-Wallis test results for cluster difference

	2 Cluster MLM	2 Cluster WLS	3 Cluster MLM	3 Cluster WLS
p Value for cluster difference	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
chi-square test statistic	106389	241900	446727	1048837

To see if two clusters differ from each other, we apply a Wilcoxon-test by rank. We report statistical significant p-values for all cluster pairs. The p-value has been adjusted by Bonferroni-Holm to avoid multicollinearity for pairwise test of difference between the clusters.

4.4.1. SEM MLM Estimation for two k-means clusters

We first apply the MLM SEM estimation to our two kmeans clusters and provide summary statistics for each cluster in Table 23. Furthermore, we provide summary statistics for the estimated latent variable which is shown by each cluster in Figure 18.

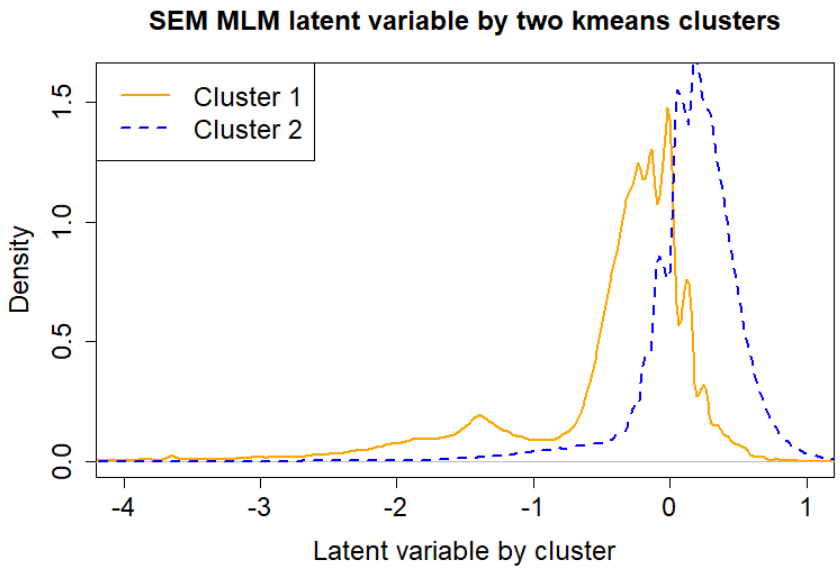

Figure 18: Density plot of latent SEM MLM variable by two clusters

The distribution of both kmeans clusters shows two quite differently distributed clusters. We analyze their difference further by the summary statistics in Table 24.

Table 24: Summary statistics of latent variable estimated by SEM MLM for two kmeans clusters

Cluster	n	mean	sd	median	min	max	skew	kurtosis
1	1446336	-0.46	0.73	-0.25	-6.50	1.18	-2.12	8.61
2	2323808	0.15	0.39	0.19	-5.62	2.03	-2.26	15.18

Table 24 shows that the two clusters differ from each other. Furthermore, we notice an unbalanced cluster size with 1,446,336 cases in the first cluster which majorly includes cases with a negative score. The second cluster includes 2,323,808 cases and shows higher range of the latent score. We analyze further what this means in a clinical context in Table 25.

Table 25: Summary statistics by variable for two kmeans clusters with SEM MLM

	Cluster	n	mean	sd	median	min	max	range	se
Apgar 5	1	1446336	9.84	0.60	10.00	6.72	12.25	10.00	0.00
	2	2323808	9.94	0.51	10.00	6.72	12.25	10.00	0.00
Apgar 10	1	1446336	11.88	0.96	12.00	0.00	12.00	12.00	0.00
	2	2323808	11.88	0.96	12.00	0.00	12.00	12.00	0.00

	Cluster	n	mean	sd	median	min	max	range	se
mother's age	1	1446336	28.71	6.02	29.00	12.00	50.00	38.00	0.01
	2	2323808	28.71	6.02	29.00	12.00	50.00	38.00	0.00
birth weight	1	1446336	6.22	0.59	6.37	4.19	6.97	2.78	0.00
	2	2323808	5.97	1.03	7.81	0.50	6.96	6.46	0.00
combined gestation	1	1446336	38.22	1.52	38.05	34.68	43.34	8.66	0.00
	2	2323808	37.54	3.02	39.00	17.00	47.00	30.00	0.00
IRT variable	1	1446336	0.14	0.75	-0.19	-0.19	5.26	5.45	0.00
	2	2323808	-0.09	0.43	-0.19	-0.19	5.26	5.45	0.00
latent variable	1	1446336	-0.46	0.73	-0.25	-6.50	1.18	7.67	0.00
	2	2323808	0.15	0.39	0.19	-5.62	2.03	7.66	0.00

By taking a closer look at the statistics of each variable by cluster of Table 25, we observe slightly higher Apgar scores, combined gestation time, and birth weight in the second cluster. However, since these values have been transformed to approximate normality a direct conclusion is not possible. Due to this transformation, the minimum Apgar 5 lays at 6.72 and the maximum is 12.25 which is not interpretable since the Apgar 5 lays between 0 and 10. The IRT variable provides more information about the health status of each cluster: cluster 1 exhibits a mean IRT score of 0.14 which indicates a higher incidence of complications. Overall, birth weight and combined gestation time show lower numbers in the second cluster as well, indicating more babies born with low-birth weight and pre-term.

4.4.2. SEM MLM Estimation for three k-means clusters

As a second step, we analyze the SEM MLM estimations choosing three kmeans clusters. First, we take a look at the latent score distribution by each cluster in Figure 19.

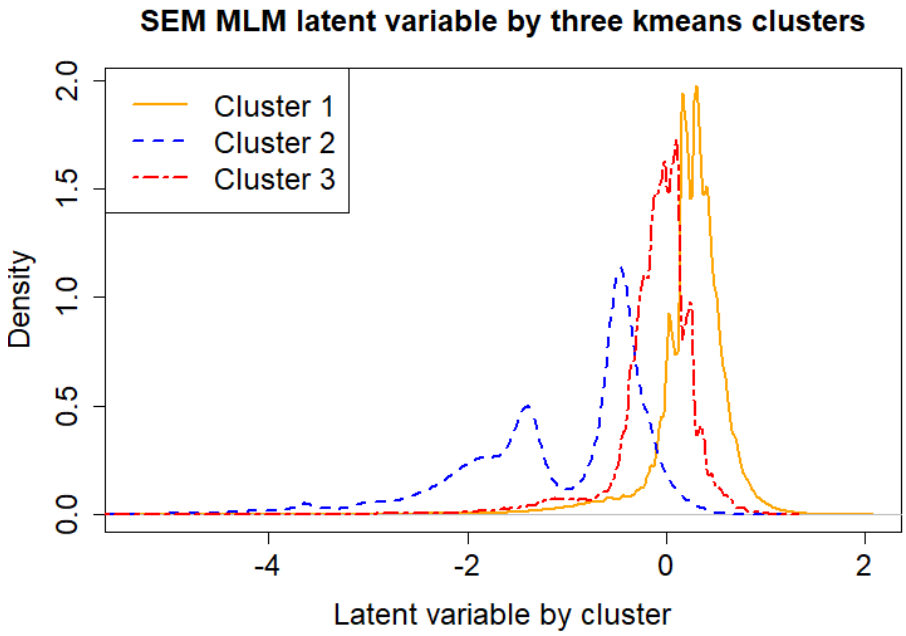

Figure 19: Density plot of latent SEM MLM variable by three clusters

Table 26: Summary statistics of latent variable estimated by SEM MLM for three kmeans clusters

cluster	n	mean	sd	median	min	max	skew	kurtosis
1	1444952	0.24	0.38	0.28	-5.31	2.03	-2.45	16.39
2	405342	-1.10	0.92	-0.70	-6.50	0.99	-1.15	4.31
3	1919850	-0.11	0.41	-0.05	-5.62	1.29	-2.46	14.71

In Figure 19, we notice that cluster 1 and cluster 2 are opposite to the previous two kmeans clusters. Here, cluster 1 shows positive scores and cluster 2 exhibits the majority of negative scores. This is supported by the summary statistics by each cluster in Table 26. We provide the summary statistics for each variable by cluster in Table 27.

Table 27: Summary statistics by variable for three kmeans clusters with SEM MLM

	Cluster	n	mean	sd	median	min	max	range	se
Apgar 5	1	1444952	9.94	0.52	10.00	6.72	12.25	5.53	0.00
	2	405342	9.64	0.75	10.00	6.72	12.25	5.53	0.00
	3	1919850	9.93	0.50	10.00	6.72	12.25	5.53	0.00
Apgar 10	1	1444952	11.97	0.43	12.00	0.00	12.00	12.00	0.00
	2	405342	11.66	1.63	12.00	0.00	12.00	0.00	0.00
	3	1919850	11.97	0.45	12.00	0.00	12.00	12.00	0.00

	Cluster	n	mean	sd	median	min	max	range	se
mother's age	1	1444952	29.40	5.58	29.00	12.00	50.00	38.00	0.00
	2	405342	28.98	6.18	29.00	12.00	50.00	38.00	0.01
	3	1919850	28.72	5.87	29.00	12.00	50.00	38.00	0.00
birth weight	1	1444952	8.34	0.58	8.20	7.61	13.79	6.18	0.00
	2	405342	5.44	0.44	5.56	4.19	5.97	1.78	0.00
	3	1919850	6.87	0.45	6.90	5.97	7.60	1.64	0.00
combined gestation	1	1444952	39.36	1.20	39.00	35.39	43.34	7.95	0.00
	2	405342	37.17	1.44	36.88	34.68	43.34	8.66	0.01
	3	1919850	38.83	1.32	39.00	35.07	43.34	8.28	0.00
IRT variable	1	1444952	-0.08	0.43	-0.19	-0.19	5.26	5.45	0.00
	2	405342	0.64	1.02	-0.19	-0.19	5.26	5.45	0.00
	3	1919850	-0.07	0.46	-0.19	-0.19	5.26	5.45	0.00
latent variable	1	1444952	0.24	0.38	0.28	-5.31	2.03	7.35	0.00
	2	405342	-1.10	0.92	-0.70	-6.50	0.99	7.79	0.00
	3	1919850	-0.11	0.41	-0.05	-5.62	1.29	6.91	0.00

We notice almost no difference for the Apgar scores across the clusters. The second cluster includes slightly more cases with lower Apgar scores indicating a not fully considered healthy baby. Additionally, mother's age varies only slightly across all clusters. A larger difference between the clusters is found in the average birth weight, combined gestation and IRT variable: cluster 1 shows the highest average birth weight and combined gestation time, as well as the lowest IRT variable on average. Therefore, cluster 1 includes the best cases of birth deliveries. Cluster 2 includes the lowest birth weight and the lowest combined gestation time on average, indicating the negatively affected birth deliveries. Furthermore, the IRT outcome for cluster 2 is with 0.63 the highest average across all clusters, stating that this cluster has the highest incidence of birth delivery complications. When we take a look at the latent variable estimation, we see the lowest average score in cluster 2 as well. However, we cannot confirm our hypothesis of three modals, because the cluster modals overlaps the three modals from the previous SEM estimation in section 4.2. Additionally, we notice apparent overlaps

between cluster 1 and 3, and cluster 2 and 3 which makes a distinction between those clusters invalid. Cluster 3 covers 2 modals from the previous SEM model in section 4.2.

4.4.3. SEM WLS Estimation for two k-means clusters

Besides SEM MLM estimation, we also test the SEM WLS estimation for our clusters. First, we apply the SEM WLS estimation to the two kmeans clusters. The graph by cluster for the latent variable is shown in Figure 20:

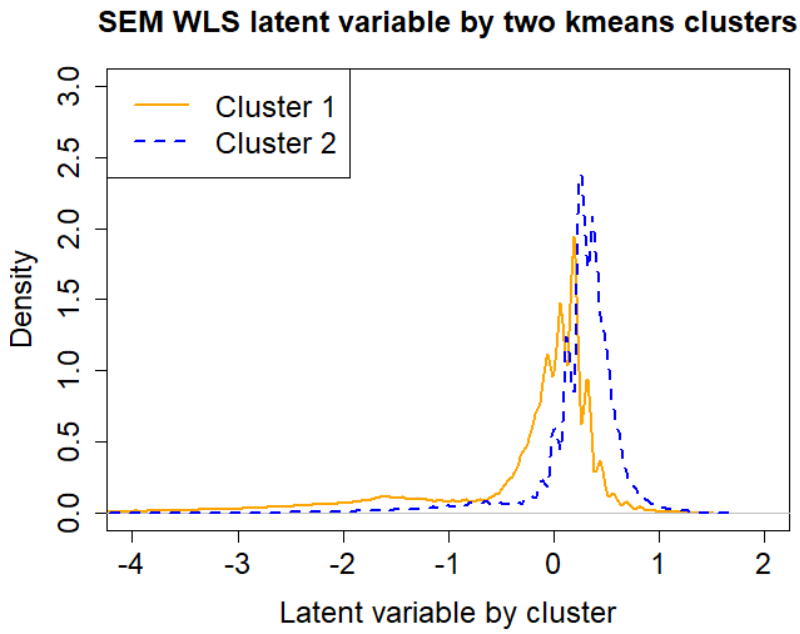

Figure 20: Density plot of latent SEM WLS variable by two clusters with focus on cluster 2

The SEM WLS clusters look quite different, we observe cluster 1 with multiple steep peaks around zero, while cluster 2 shows a much flatter curve which can be characterized as left-skewed. We present the summary statistics for the SEM WLS latent variable for two kmeans clusters in Table 28.

Table 28: Summary statistics of latent variable estimated by SEM WLS for two kmeans clusters

Cluster	n	mean	sd	median	min	max	skew	kurtosis

1	1446336	-0.37	1.00	-0.01	-8.66	1.44	-2.33	9.16
2	2323808	0.23	0.46	0.30	-6.08	1.72	-3.02	17.59

From Table 28, we notice a much higher skewness and kurtosis in cluster 1, whereas the range of the latent variable is significantly lower than cluster 2. However, both clusters show an average of zero for the latent variable. The summary statistics for each variable by kmeans clusters is presented in Table 29.

Table 29: Summary statistics by variable for two kmeans clusters with SEM WLS

	Cluster	n	mean	sd	median	min	max	range	se
Apgar 5	1	1446336	8.70	0.97	9.00	0.00	10.00	10.00	0.00
	2	2323808	8.85	0.62	9.00	0.00	10.00	10.00	0.00
Apgar 10	1	1446336	11.88	0.96	12.00	0.00	12.00	12.00	0.00
	2	2323808	11.97	0.42	12.00	0.00	12.00	12.00	0.00
mother's age	1	1446336	28.71	6.02	29.00	12.00	50.00	38.00	0.01
	2	2323808	29.19	5.66	29.00	12.00	50.00	38.00	0.00
birth weight	1	1446336	5.97	1.03	6.31	0.50	6.96	6.46	0.00
	2	2323808	7.96	0.74	7.81	6.97	18.00	11.03	0.00
combined gestation	1	1446336	37.54	3.02	38.00	17.00	47.00	30.00	0.00
	2	2323808	39.25	1.65	39.00	28.00	47.00	19.00	0.00
IRT variable	1	1446336	0.14	0.75	-0.19	-0.19	5.26	5.45	0.00
	2	2323808	-0.09	0.43	-0.19	-0.19	5.26	5.45	0.00
latent variable	1	1446336	-0.37	1.00	-0.01	-8.66	1.44	10.10	0.00
	2	2323808	0.23	0.46	0.30	-6.08	1.72	7.80	0.00

We observe higher Apgar scores on average in cluster 1, which surprisingly has a higher average age of the mother. Furthermore, cluster 1 birth deliveries have on average a longer gestation time, higher birth weight, and a lower IRT variable. From this observations we can conclude that cluster 1 birth deliveries have a slightly better quality although this is not reflected in the average by cluster of the latent variable estimation. However, the second cluster shows far more variance in the latent variable than the first cluster.

4.4.4. SEM WLS Estimation for three k-means clusters

For the SEM WLS estimation by three clusters, we provide the cluster distribution graphs in Figure 21.

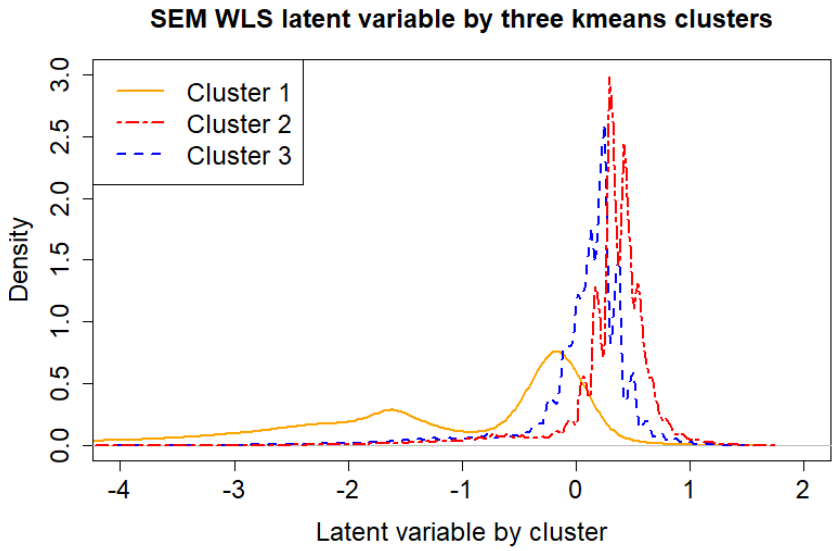

Figure 21: Density plot of latent SEM WLS variable by three clusters

From the density curves above we notice an overlap of modals across all clusters which disagrees with our assumption of three modals. We present the summary statistics of the latent SEM WLS variable for each cluster in Table 30 and the cluster characteristics in Table 31.

Table 30: Summary statistics of latent variable estimated by SEM WLS for three kmeans clusters

Cluster	n	mean	sd	median	min	max	skew	kurtosis
1	388642	-1.26	1.37	-0.75	-8.66	1.33	-1.08	3.73
2	1935671	0.05	0.51	1935671	-6.65	1.5	-2.86	15.11
3	1445831	0.28	0.46	0.35	-6.04	1.72	-3.08	17.82

Table 31: Summary statistics by variable for three kmeans clusters with SEM WLS

	Cluster	n	mean	sd	median	min	max	range	se
Apgar 5	1	388642	8.34	1.47	9.00	0.00	10.00	10.00	0.00
	2	1935671	8.85	0.64	9.00	0.00	10.00	10.00	0.00
	3	1445831	8.85	0.63	9.00	0.00	10.00	10.00	0.00
Apgar 10	1	388642	11.67	1.66	12.00	0.00	12.00	12.00	0.00
	2	1935671	11.97	0.45	12.00	0.00	12.00	12.00	0.00
	3	1445831	11.97	0.43	12.00	0.00	12.00	12.00	0.00
mother's age	1	388642	29	6.19	29.00	12.00	50.00	38.00	0.01
	2	1935671	28.71	5.87	29.00	12.00	50.00	38.00	0.01
	3	1445831	29.40	5.58	29.00	12.00	50.00	38.00	0.00
birth weight	1	388642	4.65	1.12	5.06	0.5	5.74	5.24	0.00
	2	1935671	6.83	0.49	6.89	5.74	7.60	1.86	0.00
	3	1445831	8.37	0.64	8.21	7.60	18.00	10.40	0.00
combined gestation	1	388642	35.12	3.97	36.00	17.00	47.00	30.00	0.00
	2	1935671	38.70	1.86	39.00	24.00	47.00	23.00	0.00
	3	1445831	39.39	1.61	39.00	28.00	47.00	19.00	0.00
IRT variable	1	388642	0.66	1.02	-0.19	-0.19	5.26	5.45	0.00
	2	1935671	-0.07	0.46	-0.19	-0.19	5.26	5.45	0.00
	3	1445831	-0.08	0.43	-0.19	-0.19	5.26	5.45	0.00
latent variable	1	388642	-1.26	1.37	-0.75	-8.66	1.33	9.98	0.00
	2	1935671	0.05	0.51	0.16	-6.65	1.50	8.15	0.01
	3	1445831	0.28	0.46	0.35	-6.04	1.72	7.76	0.00

To describe the latent variable in more detail, cluster 1 shows a median of 0.06, cluster 2 shows a median of 1.97, and cluster 3 shows a median of 0.02. Interesting to note about the latent variable is that its scale is reversed for SEM WLS estimation: the latent variable has the highest median in the cluster with most complications and the lowest median in the cluster which includes the healthiest cases. However, we cannot conclude that kmeans clustering proves a trimodal distribution because the cluster modals overlap.

4.4.5. DBSCAN Clustering

We apply another clustering approach, namely dbscan. This clustering method does not need the number of clusters prespecified and works on XY. We use the dbscan package in R for analysis (Hahsler M, 2019).

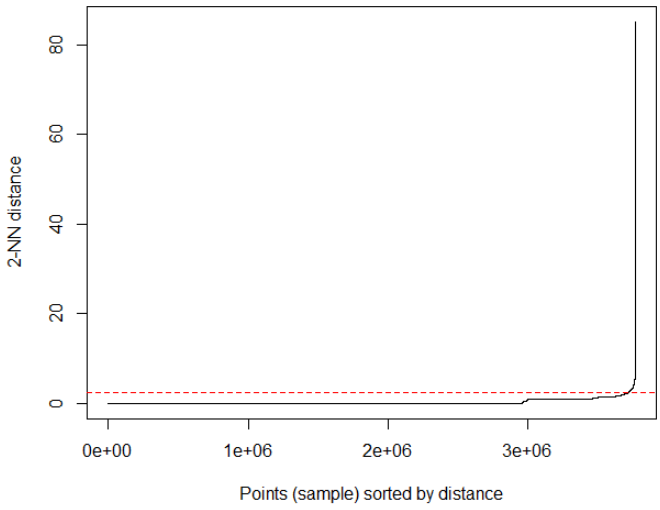

Figure 22: Screeplot for dbscan clustering to determine eps value

Similar to k-means clustering, we draw an elbow plot. Figure 22 determines the best eps value, line at 7 shows best fit. The DBSCAN Clustering algorithm does not work on more than 300,000 observations. When applying the clustering algorithm on a subsample of 300,000 cases (seed=500), the result shows 165 clusters. The clusters with amount of included cases are presented in Table 42 in the Appendix. As a second approach for dbscan clustering, we chose to apply dbscan in python using the scikit package (Pedregosa et al., 2011). The dbscan in python is resulting in 29031 clusters and 268186 noise points for the dataset using eps 7 from Figure 22. We conclude that the dbscan clustering algorithm needs further investigation in order to provide sufficient information.

4.4.6. Clustering Summary

In Table 32 we present the SEM performance comparison of the SEM clustering estimation with clusters vs the SEM MLM model from section 4.2.3. We notice a worse performance judging from a higher AIC and BIC for the MLM clustering, and higher residuals in the WLS estimation method indicating a worse model.

Table 32: Comparison of SEM performance by kmeans cluster

	SEM estimation method with kmeans clustering		Compared to SEM MLM from section 4.2.3 without clustering
Estimation method	**MLM**	**WLS**	**MLM**
n	4197370	3770144	4197370
Chi square	19.594	51.742	14.041
Robust CFI	1	1	1
Robust TLI	1	0.998	1
AIC	57610880	-	50485652
BIC	57611065	-	50485838
Robust RMSEA	0.002 [0.001 – 0.003]	0.004 [0.003 – 0.005]	0.002
Robust SRMR	0.000	0.001	0

We observe the lowest residuals and highest goodness-of-fit measures with a MLM estimator without kmeans clustering. The clustering is using SEM MLM and SEM WLS to estimate the latent variable. According to our performance overview in section 4.5.5. SEM WLS applies better to the latent variable estimation for each cluster judging from the lower residuals and better goodness-of-fit indices. This makes sense because the clustering splits the dataset into groups which are unbalanced and skewed. The WLS estimation is an asymptotically distribution-free estimation method does not require multivariate normality (Rosseel, 2012).

While the WLS estimation methods works well for the first cluster, the performance measures for the second cluster are significantly worse and the RMSEA

jumps from 0.001 for the first cluster to 0.019 for the second cluster. Applying SEM WLS for two kmeans clusters does not result in equally reliable results for each cluster. Resulting from our cluster analysis in the previous sections, we know that the second cluster represents our impaired birth deliveries. We can explain the variation by the small fraction of cases which suffer from severe complications during birth. From Figure 24, we do have at least a bimodal distribution in our underlying population. Even if it is not directly visible in Figure 24, we cannot reject the possibility of a trimodal distribution in our underlying population since all of our three clusters differed statistically significant from each other. However, the kmeans clustering is not sufficient for our dataset since the modals overlap and therefore do not create unique clusters.

4.5. Comparison of Different Methods

In this section, we compare the best models from each estimation method for the latent variable presented in the sections 4.1. – 4.2, which are namely factor analysis and SEM.

Table 33: Comparison between factor analysis model and SEM

	Factor analysis model	SEM MLM model
TLI	0.874	1
RMSEA	0.088	0.002 [0.001 – 0.003]
BIC	236041	50485838

By comparing factor analysis TLI and residuals with our SEM model in Table 33, we clearly see a better model fit by the SEM model. This is not surprising since SEM modeling can be seen as an extension to factor analysis allowing for further residual covariance specifications (Rosseel, 2012).

Since we specified paths for residual correlation of the observed variables in the SEM model, we observe a higher BIC which applies a penalty to each specified parameter (Fabozzi et al., 2014). Usually, a higher BIC indicates a worse model fit. However, a higher TLI and lower residuals suggest a better model-fit. A TLI larger than 0.9 and RMSEA smaller than 0.5 are recommended to conclude a good model fit (Xia & Yang, 2019). Therefore, we prefer the SEM model over the factor analysis model. Because of the low sample size for our HMC model, we cannot use it as a valid latent variable estimation method.

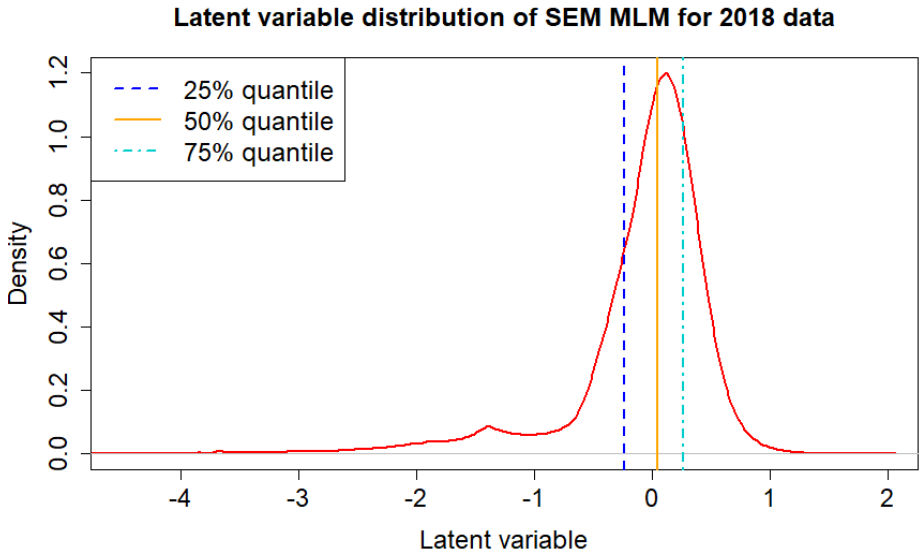

Figure 23: Percentiles for SEM MLM estimation using 2018 data

Table 34: Percentiles for SEM MLM 2018 data

Percentile	0%	25%	50%	75%	100%
Latent score	-6.45	-0.23	0.05	0.26	2.03

We conclude that the best way to model a measure for birth quality is by using a SEM MLM model. However, this requires a data transformation beforehand to approximate normality. We investigate different patients and their latent score for birth delivery quality. We identify 6 patients with fair, severe, and no problems during birth

delivery. These patients are characterized in Table 35.

Table 35: Characteristics of 6 different birth delivery cases

Case	Apgar 5	Apgar 10	Mother's age	Birth weight (grams)	Combined gestation	Maternal complications	Abnormal conditions of the newborn	Congenital abnorm-alities
1	9	88	26	4082	40	none	none	none
2	10	88	31	3147	40	none	none	none
3	5	8	39	3013	38	none	Immediate Ventilation, admission to NICU	none
4	3	5	23	3910	39	Maternal transfusion, admission to intensive care unit	Immediate ventilation, Admission to NICU, Antibiotics	none
5	0	3	25	1162	28	none	Immediate ventilation, Admission to NICU	none
6	0	4	253	595	23	none	Immediate ventilation, Admission to NICU, Surfactant, Antibiotics	none
7	9	88	32	3119	37	none	none	none
8	9	88	25	2959	39	none	none	none
9	9	88	30	2855	40	none	none	none
10	8	88	27	3620	40	none	none	none
11	9	88	33	3800	39	none	none	none
12	9	88	35	4110	36	none	Admission to NICU	none

Case 1 and 2 describe the healthier deliveries, with good Apgar scores, no complications and full-term babies. Case 3 and 4 characterize fair deliveries, with lower Apgar scores and maternal complications and/ or abnormal conditions of the newborn. Case 5 and 6 represent cases with successful CPR and pre-term birth deliveries. Case 7 – 12 show cases retrospective from Table 36, representing the 10%, 25%, 40%, 50%, 60%, and 75% percentiles of birth delivery outcomes. We notice that these cases have minimal

variation in their presented characteristics. Case 7 and 12 show a shorter gestation time but a healthy-weight baby. However, this delivery is ranked lower compared to cases 8-12 in Table 36. Cases 8-11 mainly differ in the birth weight where a heavier baby corresponds to a higher percentile. However, the underlying logic of the model that heavier babies result in better health is not true. In fact, babies with a birth weight of 10 lbs or above are considered too large. Too large babies are associated with low blood sugar, increased obesity, diabetes and metabolic syndrome (Rettner, 2013).

Since we transform the data to approximate normality, we show the same cases with their transformed values, IRT outcome, and final latent score (Health score) in Table 36.

Table 36: Outcome and data transformation of previous picked 6 different birth delivery cases sorted by Health score

Case	Apgar 5	Apgar 10	Mother's age	Birth weight (lbs)	Combined gestation	IRT outcome	Health score	Percentile for Health score
6	6.721	4	25	4.341	35.001	3.05	-4.882	0.01
5	6.721	3	25	4.629	35.391	2.041	-4.409	0.05
4	7.332	5	23	8.593	39	4.576	-4.125	0.09
3	7.876	8	39	6.662	38.051	2.041	-2.578	0.93
12	10.003	12	35	9	36.88	1.223	-0.683	10.01
7	9.186	12	32	6.881	37.356	-0.195	-0.222	25.51
8	10.003	12	25	6.556	39	-0.195	-0.05	39.58
9	10.003	12	30	6.36	39.959	-0.195	0.05	50.52
10	9.186	12	27	7.977	39.959	-0.195	0.135	60.51
11	10.003	12	33	8.363	39	-0.195	0.267	75.62
1	10.003	12	26	8.944	39.959	-0.195	0.503	92.78
2	12.254	12	31	6.941	39.959	-0.195	0.701	97.92

We notice that case 1 and 2 get a positive health score and show a quality above 90% of all cases as shown in Table 34. This assures our assumptions of a good birth delivery quality with no complications and healthy born babies. Case 3 – 5 showed complications in Table 35 which is reflected in the IRT outcome. Furthermore, their

health scores are below the 1% percentile indicating heavily impacted birth deliveries. The two pre-term and low-birth weight cases with successful CPR – case 5 and 6 – show a Health score below the 1% percentile. Case 6 shows the lightest baby in the presented cases which requires Immediate ventilation, Admission to NICU, Surfactant, and Antibiotics. This complications are reflected in the overall lowest Health score out of the presented 6 cases. Overall, this outcome matches our previous assumptions when picking those cases. A lower the health score requires more medical attention, maternal care and/ or neonatal a case receives. To test if our SEM MLM model is correctly specified, we verify the model with the birth delivery dataset from 2017.

4.6. Model Validation Using 2017 Birth Delivery Data

To verify our SEM MLM model, we apply our model to the 2017 data. The dataset includes more than 99% of all live births of birth deliveries in the United States in 2017 from citizens and non-citizens.. As our 2018 dataset, it is retrieved from the same source, the National Center for Health Statistics (NCHS) provided through the Center of Disease Control (CDC). The dataset includes 3,864,754 birth deliveries, where we move forward with 3,834,362 after data cleaning. Thus, we have a total of 30,392 missing values (0.79%). In total, we have 43,605 birth deliveries with an Apgar 10 score below 7 in 2017. The descriptive statistics for 2017 presented in Table 37 show almost the same distribution as the 2018 dataset (Table 33).

Table 37: Descriptive statistics of numerical variables for birth delivery data 2017

Variable	mean	std	skewness	kurtosis
Birth weight (g)	3264.68	588.03	-0.84	5.67
Mother's age	28.84	5.81	0.1	2.52
Combined gestation	38.62	2.44	-1.89	12.41
Apgar score after 5 minutes	8.8	0.79	-5.35	42.85
Apgar score after 10 minutes without non measured	6.02	2.66	-0.88	2.66
Apgar score after 10 minutes	87.07	8.7	-9.23	86.33

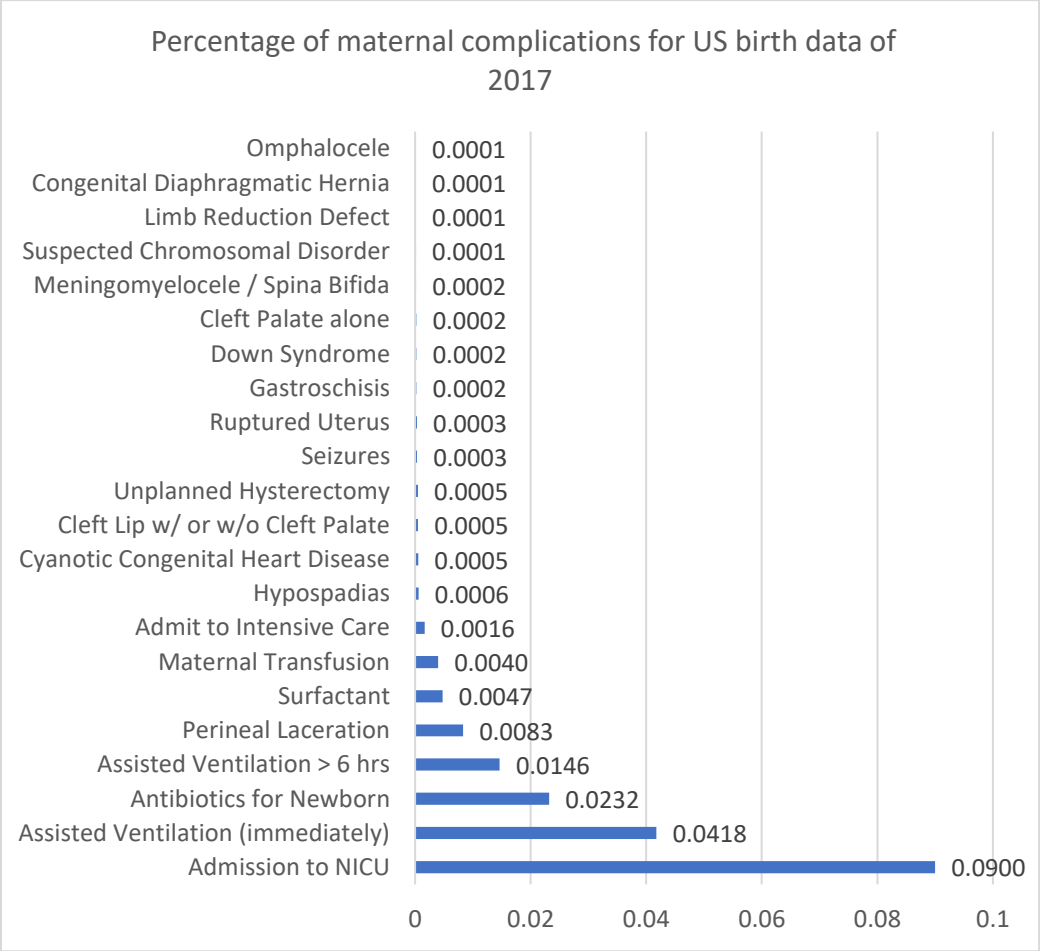

Figure 24: Frequency chart of binary variables for US births 2017

In Figure 24, we provide an overview over the binary complications in 2017. Compared to the 2018 dataset, we see very similar incidents of birth related

complications in the 2017 dataset. Admission to NICU is the most frequent occurrence in 2017 as well with 9%. We will move forward with applying the same normalization methods as for our 2018 data.

Since the datasets 2018 and 2017 are very similar, we follow the same normalization steps as described in chapter III. We apply MLM and WLS estimation for our SEM model, to see if MLM remains to outperform WLS. We present the SEM model on 2017 data in comparison to our previous 2018 dataset in Table 38.

Table 38: SEM model comparison between 2018 and 2017 data

	SEM WLS model on transformed 2018 data	**SEM MLM model on transformed 2018 data**	**SEM WLS model on transformed 2017 data**	**SEM MLM model on transformed 2017 data**
Chi square	526.044	14.041	311.236	1.15
degrees of freedom	1	1	1	1
p value	0	0	0	0.284
Robust CFI	0.999	1	0.999	1
Robust TLI	0.991	1	0.994	1
Robust RMSEA	0.012	0.002	0.009	0
RMSEA lower bound CI*	0.011	0.001	0.008	0
RMSEA upper bound CI*	0.013	0.003	0.009	0.001
Robust SRMR	0.003	0	0.002	0

*90% CI

Table 38 shows a comparison of SEM performance for the 2018 and 2017 dataset. We clearly see that our model works even better for the 2017 dataset. The SEM MLM shows less residuals and a TLI of one, indicating a good model. Furthermore, we see that the null hypothesis, our hypothesized model fits the data, is not being rejected. This shows the robustness of SEM MLM methods on the birth data, considering that the test statistic is highly sensitive for large sample sizes. We take a closer look at the distribution of the estimated birth delivery quality for 2017 (Figure 25 & Table 39).

Table 39: Latent variable distribution for 2017 data

Mean		sd	min	max	skewness	kurtosis
-0.08	0.63	-6.66	2.27		-2.35	8.8

From Figure 25 and Table 39 we notice that the results are very similar to our 2018 dataset. We see a trimodal distribution where the majority of cases varies tightly around zero. The second modal in our distribution is centered around -1.5 with a heavy left tail. This is reflected in the summary statistics for the distribution as well, it reflects our negatively skewed distribution with kurtosis. An almost unnoticeable third modal of cases peaks around -3.8. This explains our third cluster with statistical significance and should be investigated in further studies.

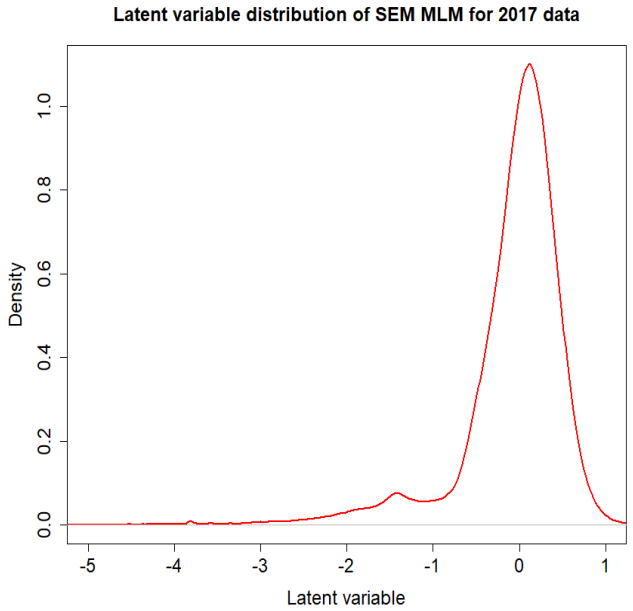

Figure 25: Density plot of latent variable estimated by SEM MLM for 2017

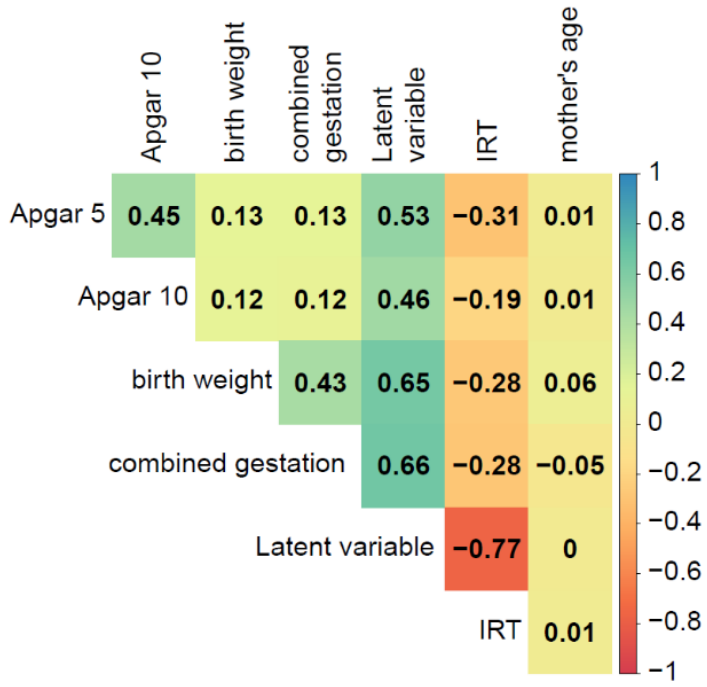

Figure 26: Heatmap for correlation of SEM MLM model variables for 2017

In Figure 26, we present a heatmap to showcase the correlation between the observed variables and our latent variable. The strongest correlation is a negative correlation between the IRT outcome, representing a strong linear association between the occurred complications and the birth delivery quality. This is because complications which likely require emergency care and directly affect mother's and baby's well-being. Our latent variable is also moderately correlated with combined gestation time, and the baby's birth weight and its Apgar 5 score. This result is similar as well to our 2018 model. Different from our 2018 dataset, we notice a stronger positive correlation between Apgar 5 and our latent variable. It is negative because we applied an inverse log transformation, hence a score of zero represents a good clinical status for the baby after 10 minutes. Compared to our 2018 dataset, we observe a much stronger correlation between mother's age and birth delivery complications.

5: CONCLUSION

In this paper, a measure for birth delivery quality was created featuring mother and child as one entity to address the lack of quality measures in the area of birth delivery. The proposed methodology is provided as a potential solution to fill the gap regarding the lack of quality measures in maternal care. The score quantitates and identifies the indirect measurable health status of the mother and newborn and provides a solid estimate for the level of required medical care. An applicable clinical method is provided to quantify the well-being of the mother and newborn. Percentiles to identify different levels of birth delivery outcomes are also provided.

Overall, a variety of models were applied: factor analysis, SEM using MLM and WLS estimation, HMC, and clustering, in order to estimate the hidden birth delivery quality. Since variables such as birth weight and gestation time are correlated, SEM proved to be the best estimation methods allowing residual covariance between the observed variables to be specified. The clustering methods k-means and dbscan did not provide sufficient results and require future study.

The estimated latent variable aimed to establish a measure to quantify birth delivery. The SEM MLM estimation model proved to be the best method for the calculating the health score. Before applying the model, data preprocessing was necessary. The binary complications: Maternal Transfusion, Ruptured Uterus, Unplanned Hysterectomy, Admit to Intensive Care, Admission to NICU, Assisted Ventilation (immediately), Surfactant, Antibiotics for Newborn, Cyanotic Congenital Heart Disease, and Cleft Lip with or without Cleft Palate were transformed into a continuous score using

a nonparametric IRT model. Then, the skewness of the data was mitigated by creating an artificial left tail of the IRT distribution. This was followed by approximating a normal distribution for the other numerical model measures: Apgar 5, birth weight, and gestation time using the R function Gaussianize from the LambertW package (Georg, 2020). The value of Apgar 10 that was not measured was replaced with the value 12 for further analysis. Lastly, the birth weight was transformed from grams to pounds.

In conclusion, a birth delivery quality measure using SEM MLM estimation resulting in a trimodal distribution was created where the third much smaller peak represents the lower quality of birth deliveries. A negative score would require medical attention and potential longer hospital stays. Furthermore, a Health score below the 25% was shown to represent severely impacted cases requiring further medical care and longer hospital stays.

The score represents an easy-to-understand measure for non-healthcare professionals with limited medical knowledge (terms and structures) of birth delivery. Strategic management can use this measure to avoid emergency bottlenecks in birth delivery areas of care. Furthermore, this measure can used to adjust hospital planning according to the needs of mothers and newborns. Considering a wider perspective, politicians can utilize this measure for policy-making and arranging long-term improvements in the birth delivery process.

While the Apgar score only takes the newborn's situation into account, this newly created score gives a comprehensive measure for the overall clinical condition of mother and child after birth delivery. It can be applied as a standardized measure for individual birth delivery quality in hospitals and allows direct comparison of birth delivery quality

between different hospitals.

The Apgar score is a score assigned by an individual doctor based on the baby's first impression and appearance. This results in a more subjective score which may vary based on an individual physician's judgement. Compared to the Apgar score, the latent score is a more comprehensive score because it includes the use of objective data: occurred complications during birth delivery, birth weight, and gestation time, in its calculation.

Siddiqui et al. (2017) uncovered a large variation of Apgar scores across European countries based on different assessments used by individual nations. Thus, model scores can be compared nationally but not across borders. Even though this model cannot be applied across borders, the methodology can be applied by individual nations: if a country has similar birth delivery measures, the proposed models (factor analysis, SEM, and clustering) can be applied to the data and the best overall model chosen based on model-fit and residual criteria, as proposed in this study. Hence, this study created a general applicable method of variable selection based on model-fit criteria that can be applied on national level in multiple countries. In this study, the developed model works in the United States, but the methodology can be applied to similar data from other countries.

A limitation of the study is the lack of different years in the model. The model was applied only to birth delivery data between 2017 and 2018. The model can be extended further as a longitudinal model to discover trends in birth delivery quality. Furthermore, the applicability to other countries was not tested. This study focused on US birth deliveries. Other countries may have different birth delivery characteristics or may

not record certain parameters such as Apgar scores. Additionally, the current model assumes that heavier babies result in better health is not true. In fact, babies with a birth weight of 10 lbs or above are considered too large and associated with low blood sugar, increased obesity, diabetes and metabolic syndrome (Rettner, 2013). To match this clinical background, the current model variable birth weight needs to be adjusted in future research. Lastly, the dataset only includes vital birth records. Mothers with a stillbirth may still have severe complications such as a ruptured uterus during birth so that the baby dies. These cases were not included in the dataset. The score only measures the timepoint immediately after birth delivery, and does not include potential complications during the pregnancy phases. The most accurate method to estimate the quality of birth deliveries is SEM using an MLM estimation. However, further investigation is needed to analyze the background and reason for those three quality categories. Further research for different clustering methods is necessary which could result in easier-to-understand groups of patients with group-specific characteristics such as pre-term birth deliveries, low birth weight, and classification of complications.

APPENDIX SECTION

Table 40: PCA vectors for 95% variance

	Explained variance	Maternal Transfusion	Perineal Laceration	Ruptured Uterus	Unplanned Hysterectomy	Admit to Intensive Care	Assisted Ventilation immediately	Assisted Ventilation > 6 hrs	Admission to NICU	Surfactant	Antibiotics for Newborn	Seizures
1	0.560	0.010	-0.001	0.001	0.002	0.008	0.437	0.215	0.841	0.063	0.227	0.004
2	0.185	0.008	0.005	0.001	0.001	0.002	0.777	0.305	-0.526	0.074	0.143	0.003
3	0.090	0.008	0.008	0.000	0.000	-0.001	-0.321	0.143	-0.126	0.098	0.923	0.005
4	0.046	0.009	0.498	0.001	0.000	-0.001	0.276	-0.767	0.005	-0.185	0.230	-0.002
5	0.045	0.021	0.867	0.000	0.002	0.004	-0.159	0.439	0.003	0.101	-0.142	0.001
6	0.023	0.974	-0.022	0.018	0.078	0.197	-0.007	-0.025	-0.004	0.060	-0.014	0.002
7	0.020	-0.066	0.005	-0.001	-0.004	-0.003	0.014	-0.243	-0.001	0.966	-0.059	-0.002

	Anencephaly	Meningom-yelocele / Spina Bifida	Cyanotic Congenital Heart Disease	Congenital Diaphrag-matic Hernia	Omphalocele	Gastroschisis	Limb Reduction Defect	Cleft Lip w/ or w/o Cleft Palate	Cleft Palate alone	Down Syndrome	Suspected Chromos-omal Disorder	Hypospadias
1	0.000	0.001	0.005	0.001	0.001	0.002	0.000	0.001	0.001	0.001	0.001	0.001
2	0.000	-0.001	-0.001	0.001	0.000	-0.001	0.000	-0.001	0.000	0.000	0.000	0.000
3	0.000	0.000	-0.001	0.000	0.000	0.001	0.000	-0.001	0.000	0.000	0.000	0.001
4	0.000	0.000	-0.003	-0.001	0.000	0.000	0.000	0.000	0.000	-0.001	-0.001	-0.001
5	0.000	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	-0.001	0.000	0.000	0.000	-0.001	0.000	0.000
7	0.000	0.000	-0.002	-0.001	0.000	-0.002	0.000	0.000	0.000	-0.002	0.000	0.000

Table 41: PCA vectors for 99% variance part I

	Explained variance	Maternal Transfusion	Perineal Laceration	Ruptured Uterus	Unplanned Hysterectomy	Admit to Intensive Care	Assisted Ventilation (imme-diately)	Assisted Ventilation > 6 hrs	Admission to NICU	Surfactant	Antibiotics for Newborn	Seizures
		mm_mtr	mm_plac	mm_rupt	mm_uhyst	mm_aicu	ab_aven1	ab_aven6	ab_nicu	ab_surf	ab_anti	ab_seiz
1	0.560	0.010	-0.001	0.001	0.002	0.008	0.437	0.215	0.841	0.063	0.227	0.004
2	0.185	0.008	0.005	0.001	0.001	0.002	0.777	0.305	-0.526	0.074	0.143	0.003
3	0.090	0.008	0.008	0.000	0.000	-0.001	-0.321	0.143	-0.126	0.098	0.923	0.005
4	0.046	0.009	0.498	0.001	0.000	-0.001	0.276	-0.767	0.005	-0.185	0.230	-0.002
5	0.045	0.021	0.867	0.000	0.002	0.004	-0.159	0.439	0.003	0.101	-0.142	0.001
6	0.023	0.974	-0.022	0.018	0.078	0.197	-0.007	-0.025	-0.004	0.060	-0.014	0.002
7	0.020	-0.066	0.005	-0.001	-0.004	-0.003	0.014	-0.243	-0.001	0.966	-0.059	-0.002
8	0.008	-0.207	0.002	0.024	0.143	0.968	-0.003	0.000	-0.005	-0.010	0.002	0.002
9	0.003	0.000	0.000	0.001	0.000	0.000	0.000	-0.001	-0.001	-0.001	-0.001	-0.001
10	0.003	-0.001	0.000	0.002	0.008	0.000	-0.001	-0.004	-0.005	0.001	0.001	0.006
11	0.003	0.000	0.000	0.000	-0.005	0.001	0.000	0.001	0.000	0.000	0.000	0.001
12	0.002	-0.049	0.000	0.254	0.953	-0.157	0.000	0.000	0.000	0.000	0.001	0.015
13	0.002	-0.001	-0.001	0.805	-0.221	0.011	-0.002	-0.002	-0.001	0.000	-0.003	0.550

Table 42: PCA vectors for 99% variance part II

	Anencephaly	Meningom-yelocele / Spina Bifida	Cyanotic Congenital Heart Disease	Congenital Diaphragmatic Hernia	Omphalocele	Gastroschisis	Limb Reduction Defect	Cleft Lip w/ or w/o Cleft Palate	Cleft Palate alone	Down Syndrome	Suspected Chromosomal Disorder	Hypospadias
	ca_anen	ca_mnsb	ca_cchd	ca_cdh	ca_omph	ca_gast	ca_limb	ca_cleft	ca_clpal	ca_downs	ca_disor	ca_hypo
1	0.000	0.001	0.005	0.001	0.001	0.002	0.000	0.001	0.001	0.001	0.001	0.001
2	0.000	-0.001	-0.001	0.001	0.000	-0.001	0.000	-0.001	0.000	0.000	0.000	0.000
3	0.000	0.000	-0.001	0.000	0.000	0.001	0.000	-0.001	0.000	0.000	0.000	0.001
4	0.000	0.000	-0.003	-0.001	0.000	0.000	0.000	0.000	0.000	-0.001	-0.001	-0.001
5	0.000	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	-0.001	0.000	0.000	0.000	-0.001	0.000	0.000
7	0.000	0.000	-0.002	-0.001	0.000	-0.002	0.000	0.000	0.000	-0.002	0.000	0.000
8	0.000	0.000	-0.002	0.000	0.000	-0.001	0.000	0.000	0.000	-0.001	0.000	0.000
9	0.001	0.001	0.046	0.003	0.001	-0.001	0.004	0.061	0.011	0.004	0.008	0.997
10	0.005	0.006	0.944	0.012	0.013	0.001	0.008	0.311	0.042	0.053	0.043	-0.063
11	0.005	0.000	-0.318	-0.001	0.000	0.002	-0.001	0.940	0.117	-0.015	0.011	-0.044
12	0.000	0.002	-0.009	0.000	0.000	-0.001	0.001	0.002	-0.001	0.000	-0.001	0.000
13	0.006	0.002	-0.002	0.000	0.000	-0.003	0.001	-0.002	-0.002	-0.004	0.004	0.001

Table 43: mPCA vectors 99% variance Matrix of Squared loadings

Variable	Dimension																										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
apgar5	.40	.03	.38	.03	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
apgar10	.27	.05	.49	.04	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.01
mager	.00	.00	.00	.00	.00	.01	.13	.05	.36	.18	.09	.00	.03	.10	.00	.00	.00	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00
dbwt	.29	.00	.00	.42	.01	.00	.00	.00	.00	.01	.00	.00	.00	.00	.02	.01	.00	.00	.00	.00	.00	.00	.00	.00	.03	.20	.00
combgest	.35	.00	.00	.36	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00	.01	.00	.00	.02	.23	.00
mm_mtr	.01	.34	.05	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.01	.00	.00	.00	.57	.00	.00	.00	.00	.00	.00
mm_plac	.00	.00	.00	.03	.00	.00	.00	.02	.15	.03	.41	.03	.05	.13	.13	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
mm_rupt	.00	.08	.00	.00	.00	.00	.00	.01	.02	.00	.02	.00	.02	.04	.45	.10	.24	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00
mm_uhyst	.00	.43	.05	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.15	.00	.35	.00	.00	.00	.00
mm_aicu	.01	.42	.04	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.02	.00	.01	.00	.00	.00	.09	.00	.39	.00	.00	.00	.00
ab_aven1	.48	.01	.03	.07	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.04	.00	.17	.00	.00	.17
ab_aven6	.44	.02	.12	.07	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.10	.04	.00	.19
ab_nicu	.46	.01	.02	.02	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.13	.00	.02	.31	.00	.02

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
ab_surf	.24	.01	.07	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.02	.01	.00	.00	.00	.54	.00	.01	.05	.00	.01
ab_anti	.28	.02	.11	.06	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.04	.00	.39	.08	.00	.01
ab_seiz	.01	.00	.00	.08	.00	.00	.00	.13	.01	.03	.10	.00	.01	.00	.15	.17	.29	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00
ca_anen	.00	.00	.02	.00	.01	.00	.02	.23	.00	.12	.02	.00	.07	.40	.01	.01	.08	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00
ca_mnsb	.00	.00	.00	.00	.03	.03	.08	.10	.06	.04	.11	.27	.06	.10	.00	.10	.04	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
ca_cchd	.01	.00	.00	.02	.15	.12	.03	.01	.03	.00	.02	.00	.01	.00	.00	.00	.01	.58	.00	.00	.00	.01	.00	.00	.00	.00	.00
ca_cdh	.00	.00	.00	.01	.01	.01	.00	.27	.04	.02	.02	.01	.33	.07	.10	.00	.04	.05	.00	.00	.00	.00	.00	.00	.00	.00	.00
ca_omph	.00	.00	.00	.00	.09	.08	.24	.00	.02	.05	.01	.00	.00	.00	.00	.18	.06	.01	.25	.00	.00	.00	.00	.00	.00	.00	.00
ca_gast	.00	.00	.00	.00	.01	.01	.37	.00	.00	.05	.00	.15	.05	.02	.00	.13	.06	.00	.11	.00	.00	.01	.00	.00	.00	.00	.00
ca_limb	.00	.00	.00	.00	.02	.01	.00	.12	.09	.11	.10	.14	.33	.04	.02	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
ca_cleft	.00	.00	.00	.00	.23	.31	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.45	.00	.00	.00	.00	.00	.00	.00
ca_clpal	.00	.00	.00	.00	.19	.34	.00	.00	.00	.01	.00	.00	.00	.00	.00	.01	.01	.00	.02	.41	.00	.00	.00	.00	.00	.00	.00
ca_downs	.00	.00	.00	.00	.11	.11	.15	.01	.07	.00	.00	.03	.01	.00	.00	.14	.06	.13	.16	.01	.00	.00	.00	.00	.00	.00	.00
ca_disor	.00	.00	.01	.01	.24	.06	.02	.01	.01	.01	.00	.00	.00	.02	.00	.10	.02	.12	.36	.01	.00	.00	.00	.00	.00	.00	.00

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
ca_hypo	.00	.00	.00	.00	.00	.00	.00	.04	.15	.34	.08	.35	.00	.03	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Table 44: Variable names and their actual meaning

apgar5	apgar10	mager	dbwt	combgest	mm_mtr	mm_plac	mm_rupt	mm_uhyst	mm_aicu
Apgar 5	Apgar 10	mother's age	Birth weight	combined gestation	maternal transfusion	Perineal Laceration	Ruptured Uterus	Unplanned Hysterectomy	Admit to Intensive Care

ab_aven1	ab_aven6	ab_nicu	ab_surf	ab_anti	ab_seiz	ca_anen	ca_mnsb	ca_cchd	ca_cdh
Assisted Ventilation (immediately)	Assisted Ventilation > 6 hrs	Admission to NICU	Surfactant	Antibiotics for Newborn	Seizures	Anen-cephaly	Meningomyelocele / Spina Bifida	Cyanotic Congenital Heart Disease	Congenital Diaphragmatic Hernia

ca_gast	ca_limb	ca_cleft	ca_clpal	ca_downs	ca_disor	ca_hypo
Gastroschisis	Limb Reduction Defect	Cleft Palate alone	Cleft Lip w/ or w/o Cleft Palate	Down Syndrome	Suspected Chromosomal Disorder	Hypos-padias

Table 45: dbscan clustering clusters with number of cases

Cluster number	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of cases	1469	295647	103	11	15	7	293	147	19	56	26	24	27	36	14	59
Cluster number	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Number of cases	83	19	38	22	9	14	15	67	21	13	26	31	26	12	10	7
Cluster number	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
Number of cases	28	38	11	48	15	43	5	15	33	25	9	11	29	5	27	8
Cluster number	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
Number of cases	40	14	9	8	16	20	25	9	53	27	6	16	7	5	14	6
Cluster number	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
Number of cases	11	17	16	24	17	11	13	10	13	17	11	45	12	9	5	33
Cluster number	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
Number of cases	14	12	20	21	51	10	9	9	5	8	6	19	9	27	10	8
Cluster number	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111
Number of cases	7	6	13	22	7	5	20	5	5	12	21	19	9	6	7	8
Cluster number	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127
Number of cases	5	8	9	5	7	8	4	11	12	5	6	10	4	11	5	5
Cluster number	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
Number of cases	5	4	10	5	5	13	6	6	5	5	7	7	7	5	10	6
Cluster number	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
Number of cases	5	6	6	6	4	6	5	6	5	4	5	5	5	5	6	5

Cluster number	160	161	162	163	164	165
Number of cases	6	5	5	4	5	5

REFERENCES

Adnan, R., & Thanoon, Y. T. (2015). Bayesian analysis of multiple group nonlinear structural equation models with ordered categorical and dichotomous variables:a survey.

American College of Obstetricians and Gynecologists. (2015). Committee Opinion No. 644: The Apgar Score. *Obstet Gynecol, 126*(4), e52-e55. https://doi.org/10.1097/aog.0000000000001108

Amiri, L., Khazaei, M., & Ganjali, M. (2018). A mixture latent variable model for modeling mixed data in heterogeneous populations and its applications. *AStA Advances in Statistical Analysis, 102*(1), 95-115. https://doi.org/10.1007/s10182-017-0294-3

Asparouhov, T., & Muthén, B. (2016). Structural Equation Models and Mixture Models With Continuous Nonnormal Skewed Distributions. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(1), 1-19. https://doi.org/10.1080/10705511.2014.947375

Azzalini, A., & Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika, 83*(4), 715-726. https://doi.org/10.1093/biomet/83.4.715

Bauer, D. J., & Curran, P. J. (2004, Mar). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychol Methods, 9*(1), 3-29. https://doi.org/10.1037/1082-989x.9.1.3

Bhat, C. R. (2015). A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B: Methodological, 79*, 50-77. https://doi.org/https://doi.org/10.1016/j.trb.2015.05.017

Bhat, C. R., Pinjari, A. R., Dubey, S. K., & Hamdi, A. S. (2016). On accommodating spatial interactions in a Generalized Heterogeneous Data Model (GHDM) of mixed types of dependent variables. *Transportation Research Part B: Methodological, 94*, 240-263. https://doi.org/https://doi.org/10.1016/j.trb.2016.09.002

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1). https://doi.org/10.18637/jss.v080.i01

Cai, J.-H., Song, X.-Y., Lam, K.-H., & Ip, E. H.-S. (2011). A mixture of generalized latent variable models for mixed mode and heterogeneous data. *Computational Statistics & Data Analysis, 55*(11), 2889-2907. https://doi.org/https://doi.org/10.1016/j.csda.2011.05.011

Center for Disease Control and Prevention. (2020a). *Facts about Anencephaly*. Retrieved 1/6/2021 from https://www.cdc.gov/ncbddd/birthdefects/anencephaly.html

Center for Disease Control and Prevention. (2020b). *Severe Maternal Morbidity*. Retrieved 1/6/2021 from https://www.cdc.gov/reproductivehealth/maternalinfanthealth/severematernalmorbidity.html

Centers for Disease Control and Prevention. (2020). *Cleft Lip / Cleft Palate*. Retrieved 05/17/2021 from https://www.cdc.gov/ncbddd/birthdefects/cleftlip.html

Chakkarapani, A. A., Adappa, R., Mohammad Ali, S. K., Gupta, S., Soni, N. B., Chicoine, L., & Hummler, H. D. (2020). "Current concepts of mechanical ventilation in neonates" - Part 1: Basics. *International journal of pediatrics & adolescent medicine, 7*(1), 13-18. https://doi.org/10.1016/j.ijpam.2020.03.003

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and Absolute Fit Evaluation in Cognitive Diagnosis Modeling. *Journal of Educational Measurement, 50*(2), 123-140. https://doi.org/https://doi.org/10.1111/j.1745-3984.2012.00185.x

Cirino, E. (2017). *Pregnancy Complications: Uterine Rupture*. Retrieved 05/16/2021 from https://www.healthline.com/health/pregnancy/complications-uterine-rupture

Civek, B. C., & Kozat, S. S. (2017). Efficient Implementation of Newton-Raphson Methods for Sequential Data Prediction. *IEEE Transactions on Knowledge and Data Engineering, 29*(12), 2786-2791. https://doi.org/10.1109/TKDE.2017.2754380

Collier, A.-R. Y., & Molina, R. L. (2019). Maternal Mortality in the United States: Updates on Trends, Causes, and Solutions. *NeoReviews, 20*(10), e561-e574. https://doi.org/10.1542/neo.20-10-e561

Contreras-Reyes, J. E., & Arellano-Valle, R. B. (2013). Growth estimates of cardinalfish (Epigonus crassicaudus) based on scale mixtures of skew-normal distributions. *Fisheries Research, 147*, 137-144. https://doi.org/https://doi.org/10.1016/j.fishres.2013.05.002

Cui, R. (2019). *Causal Discovery from Mixed Data using Gaussian Copula Models* [Dissertation, Radboud University Nijmegen]. Nijmegen.

de Leon, A. R., & Carriègre, K. C. (2007). General mixed-data model: Extension of general location and grouped continuous models. *Canadian Journal of Statistics, 35*(4), 533-548. https://doi.org/https://doi.org/10.1002/cjs.5550350405

de Leon, A. R., & Carrière, K. C. (2005). A generalized Mahalanobis distance for mixed data. *Journal of Multivariate Analysis, 92*(1), 174-185. https://doi.org/https://doi.org/10.1016/j.jmva.2003.08.006

de Leon, A. R., & Chough, K. C. (2013). *Analysis of Mixed Data: Methods & Applications*. Taylor & Francis. https://books.google.com/books?id=tObrm_XrLWIC

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. *Journal of Computational and Graphical Statistics, 26*(2), 403-413. https://doi.org/10.1080/10618600.2016.1172487

Depaoli, S., & Clifton, J. P. (2015). A Bayesian Approach to Multilevel Structural Equation Modeling With Continuous and Dichotomous Outcomes. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(3), 327-351. https://doi.org/10.1080/10705511.2014.937849

Dirlik, E. M. (2019). The Comparison of Item Parameters Estimated From Parametric and Nonparametric Item Response Theory Models in Case of The Violance of Local Independence Assumption. *International Journal of Progressive Education, 15*(4), 229-240. https://doi.org/10.29329/ijpe.2019.203.17

Elmir, R., Jackson, D., Schmied, V., & Wilkes, L. "Less Feminine and Less a Woman":The Impact of Unplanned Postpartum Hysterectomy on Women. *Int J Childbirth*(1), 51-60. https://doi.org/10.1891/2156-5287.2.1.51

Fabozzi, F. J., Focardi, S. M., Rachev, S. T., & Arshanapalli, B. G. (2014). Appendix E: Model Selection Criterion: AIC and BIC. In *The Basics of Financial Econometrics* (pp. 399-403). https://doi.org/https://doi.org/10.1002/9781118856406.app5

Fahrmeir, L., & Raach, A. (2007). A Bayesian Semiparametric Latent Variable Model for Mixed Responses. *Psychometrika, 72*(3), 327. https://doi.org/10.1007/s11336-007-9010-7

Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S. R., Park, H., & Shao, C. (2016). Applications of structural equation modeling (SEM) in ecological studies: an updated review. *Ecological Processes, 5*(1), 19. https://doi.org/10.1186/s13717-016-0063-3

Feng, X.-N., Wu, H.-T., & Song, X.-Y. (2017). Bayesian Regularized Multivariate Generalized Latent Variable Models. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(3), 341-358. https://doi.org/10.1080/10705511.2016.1257353

Feng, X., Lu, B., Song, X., & Ma, S. (2019). Financial literacy and household finances: A Bayesian two-part latent variable modeling approach. *Journal of Empirical Finance, 51*, 119-137. https://doi.org/https://doi.org/10.1016/j.jempfin.2019.02.002

Ford, C. (2016). *Getting Started with Factor Analysis*. University of Virginia Library StatLab. Retrieved 05/16/2021 from https://data.library.virginia.edu/getting-started-with-factor-analysis/

Frushicheva, M. P. (2016). *k-means Clustering*. RStudio Retrieved 05/06/2021 from https://rpubs.com/violetgirl/201598

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian Regression Models. *The American Statistician, 73*(3), 307-309. https://doi.org/10.1080/00031305.2018.1549100

Georg, G. (2020). *LambertW: Probabilistic Models to Analyze and Gaussianize Heavy-Tailed, Skewed Data.* In [R package version 0.6.6.].

Glen, S. (2021a). *Communality: Definition, Examples*. StatisticsHowTo.com: Elementary Statistics for the rest of us! . Retrieved 05/19/2021 from https://www.statisticshowto.com/communality/

Glen, S. (2021b). *Kruskal Wallis H Test: Definition, Examples & Assumptions*. Statistics How To: Elementary Statistics for the rest of us! Retrieved 05/27 from https://www.statisticshowto.com/kruskal-wallis/

Glen, S. (2021c). *RMSE: Root Mean Square Error*. Statistics How To: Elementary Statistics for the rest of us! Retrieved 5/12/2021 from https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

Goerg, G. M. (2015). The Lambert Way to Gaussianize Heavy-Tailed Data with the Inverse of Tukey's Transformation as a Special Case. *The Scientific World Journal, 2015*, 909231. https://doi.org/10.1155/2015/909231

Griffin, M. M., & Steinbrecher, T. D. (2013). Chapter Four - Large-Scale Datasets in Special Education Research. In R. C. Urbano (Ed.), *International Review of Research in Developmental Disabilities* (Vol. 45, pp. 155-183). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-407760-7.00004-9

Gruhl, J., Erosheva, E. A., & Crane, P. K. (2013). A semiparametric approach to mixed outcome latent variable models: Estimating the association between cognition and regional brain volumes. *Ann. Appl. Stat., 7*(4), 2361-2383. https://doi.org/10.1214/13-AOAS675

Hahsler M, P. M., Doran D. (2019). dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software, 91*(1), 1-30. https://doi.org/10.18637/jss

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 621-638. https://doi.org/10.1080/10705511.2017.1402334

Health Affairs Blog. (2019). *Addressing The US Maternal Health Crisis: Policies Of 2020 Presidential Candidates*. Retrieved 12/15/2020 from https://www.healthaffairs.org/do/10.1377/hblog20190625.583781/full/

Hoyle, R. H. (2011). Structural Equation Modeling for Social and Personality Psychology. In. SAGE Publications Ltd. https://doi.org/10.4135/9781446287965

Jafari, N., Tabrizi, E., & Samani, E. B. (2015). Gaussian Copula Mixed Models with Non-Ignorable Missing Outcomes. *Applications & Applied Mathematics, 10*(1), 25.

Jiryaie, F., Withanage, N., Wu, B., & de Leon, A. R. (2016). Gaussian copula distributions for mixed data, with application in discrimination. *Journal of Statistical Computation and Simulation, 86*(9), 1643-1659. https://doi.org/10.1080/00949655.2015.1077386

Jöreskog, K. G. (1977). Factor analysis by least squares and maximum likelihood methods. In *Statistical Methods for Digital Computers*. Wiley, New York.

Jr, P. J. R., Diggle, P. J., Schlather, M., Bivand, R., & Ripley, B. (2020). *geoR: Analysis of Geostatistical Data.* In (Version 1.8-1) https://CRAN.R-project.org/package=geoR

Kassambara, A. (2018). *Transform Data to Normal Distribution in R* Datanovia. Retrieved 05/17/2021 from https://www.datanovia.com/en/lessons/transform-data-to-normal-distribution-in-r/

Kelava, A., & Brandt, H. (2014). A general non-linear multilevel structural equation mixture model [Methods]. *Frontiers in Psychology, 5*(748). https://doi.org/10.3389/fpsyg.2014.00748

Kenny, D. A. (2020). *Measuring Model Fit*. Retrieved 05/18/2021 from http://www.davidakenny.net/cm/fit.htm

Khatab, K. (2007). *Analysis of Childhood Diseases and Malnutrition in Developing Countries of Africa* [Dissertation, LMU Munich]. Munich. http://nbn-resolving.de/urn:nbn:de:bvb:19-77270

Koh, H. K. (2011). The ultimate measures of health. *Public health reports (Washington, D.C. : 1974), 126 Suppl 3*, 14-15. https://doi.org/10.1177/00333549111260S303

Korang, S. K., Safi, S., Gluud, C., Lausten-Thomsen, U., & Jakobsen, J. C. (2019). Antibiotic regimens for neonatal sepsis - a protocol for a systematic review with meta-analysis. *Systematic Reviews, 8*(1), 306. https://doi.org/10.1186/s13643-019-1207-1

Kozhimannil, K. B. (2014). *Better Measurement Of Maternity Care Quality*. Retrieved 05/02/2021 from https://www.healthaffairs.org/do/10.1377/hblog20140812.040717/full/

Kunihama, T. (2015). *Nonparametric Bayes Analysis of Social Science Data* [Dissertation, Duke University].

Lee, S. Y., & Shi, J. Q. (2001, Sep). Maximum likelihood estimation of two-level latent variable models with mixed continuous and polytomous data. *Biometrics, 57*(3), 787-794. https://doi.org/10.1111/j.0006-341x.2001.00787.x

Lee, S. Y., & Song, X. Y. (2004, May). Bayesian model comparison of nonlinear structural equation models with missing continuous and ordinal categorical data. *The British journal of mathematical and statistical psychology, 57*(Pt 1), 131-150. https://doi.org/10.1348/000711004849204

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature Selection: A Data Perspective. *ACM Comput. Surv., 50*(6), Article 94. https://doi.org/10.1145/3136625

Lin, T.-I., Ho, H. J., & Lee, C.-R. (2014). Flexible mixture modelling using the multivariate skew-t-normal distribution. *Statistics and Computing, 24*(4), 531-546. https://doi.org/10.1007/s11222-013-9386-4

Lin, T.-I., McLachlan, G. J., & Lee, S. X. (2016). Extending mixtures of factor models using the restricted multivariate skew-normal distribution. *Journal of Multivariate Analysis, 143*, 398-413. https://doi.org/https://doi.org/10.1016/j.jmva.2015.09.025

Lubke, G. H., & Luningham, J. (2017, Nov). Fitting latent variable mixture models. *Behaviour research and therapy, 98*, 91-102. https://doi.org/10.1016/j.brat.2017.04.003

Lukasz Komsta, F. N. (2015). *Package 'moments'*. Retrieved 1/6/2021 from https://CRAN.R-project.org/package=moments

Ma, Z., & Chen, G. (2019). Bayesian semiparametric latent variable model with DP prior for joint analysis: Implementation with nimble. *Statistical Modelling, 20*(1), 71-95. https://doi.org/10.1177/1471082X18810118

Ma, Z., & Chen, G. (2020). Bayesian joint analysis using a semiparametric latent variable model with non-ignorable missing covariates for CHNS data. *Statistical Modelling, 0*(0), 1471082X19896688. https://doi.org/10.1177/1471082x19896688

Machado, L. S. M. (2011). Emergency peripartum hysterectomy: Incidence, indications, risk factors and outcome. *North American journal of medical sciences, 3*(8), 358-361. https://doi.org/10.4297/najms.2011.358

Malik, U. (2018). *Implementing PCA in Python with Scikit-Learn*. Retrieved 1/4/2021 from https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/

Mangiafico, S. (2021). *rcompanion: Functions to Support Extension Education Program Evaluation.* In (Version 2.4.1) CRAN. https://CRAN.R-project.org/package=rcompanion

Marchese, S. (2018). *Semiparametric Regression Methods for Mixed Type Data Analysis* [Dissertation, George Mason University]. Fairfax, VA.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *2011, 42*(9), 21. https://doi.org/10.18637/jss.v042.i09

Mayo Clinic. (2019). *Vaginal tears in childbirth*. Mayo Foundation for Medical Education and Research (MFMER). Retrieved 1/6/2021 from https://www.mayoclinic.org/healthy-lifestyle/labor-and-delivery/multimedia/vaginal-tears/sls-20077129?s=5

Mayo Clinic. (n.d.). *Hypospadias*. Mayo Clinic. Retrieved 05/17/2021 from https://www.mayoclinic.org/diseases-conditions/hypospadias/symptoms-causes/syc-20355148

McDowell, I., Spasoff, R. A., & Kristjansson, B. (2004). On the classification of population health measurements. *American journal of public health, 94*(3), 388-393. https://doi.org/10.2105/ajph.94.3.388

McManus, R. W., & Nieman, M. D. (2019). Identifying the level of major power support signaled for protégés: A latent measure approach. *Journal of Peace Research, 56*(3), 364-378. https://doi.org/10.1177/0022343318808842

McParland, D., Phillips, C., Brennan, L., Roche, H. M., & Gormley, I. C. (2017). Clustering high-dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data. *Stat Med., 36*(28), 4548-4569. https://doi.org/doi:10.1002/sim.7371

Meijer, E., Kapteyn, A., & Andreyeva, T. (2011). Internationally comparable health indices. *Health Econ, 20*(5), 600-619. https://doi.org/10.1002/hec.1620

Melbourne, T. R. C. s. H. (2018). *Surfactant Administration in the NICU*. Retrieved 1/6/2021 from https://www.rch.org.au/rchcpg/hospital_clinical_guideline_index/Surfactant_Administration_in_the_NICU/

Meloun, M., & Militky, J. (2011). 4 - Statistical analysis of multivariate data. In *Statistical Data Analysis: A Practical Guide* (pp. 151-403). Woodhead Publishing, Limited.

Meyer, J. (2020). *First Steps in Structural Equation Modeling: Confirmatory Factor Analysis*. The Analysis Factor. Retrieved 05/15/2021 from https://www.theanalysisfactor.com/structural-equation-modeling-first-step-confirmatory-factor-analysis-2/

Mikulski, B. (2019). *PCA—how to choose the number of components?* Retrieved 1/4/2021 from https://www.mikulskibartosz.name/pca-how-to-choose-the-number-of-components/

Moller, A.-B., Newby, H., Hanson, C., Morgan, A., El Arifeen, S., Chou, D., Diaz, T., Say, L., Askew, I., & Moran, A. C. (2018). Measures matter: A scoping review of maternal and newborn indicators. *PLoS One, 13*(10), e0204763-e0204763. https://doi.org/10.1371/journal.pone.0204763

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution, 8*(3), 339-348. https://doi.org/https://doi.org/10.1111/2041-210X.12681

Morlini, I. (2012). A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model. *Advances in Data Analysis and Classification, 6*(1), 5-28. https://doi.org/10.1007/s11634-011-0101-z

Murray, C. J., & Frenk, J. (2008). Health metrics and evaluation: strengthening the science. *Lancet, 371*(9619), 1191-1199. https://doi.org/10.1016/s0140-6736(08)60526-7

Murray, J. S., Dunson, D. B., Carin, L., & Lucas, J. E. (2013). Bayesian Gaussian Copula Factor Models for Mixed Data. *Journal of the American Statistical Association, 108*(502), 656-665.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132. https://doi.org/10.1007/BF02294210

Muthén, B., & Satorra, A. (1995). Technical aspects of Muthén's liscomp approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika, 60*(4), 489-503. https://EconPapers.repec.org/RePEc:spr:psycho:v:60:y:1995:i:4:p:489-503

National Center for Health Statistics. (2019). *User Guide to the 2018 Natality Public Use File*. Retrieved 5/7/2021 from http://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm

Nussbaum, F., & Giesen, J. (2020). Pairwise sparse + low-rank models for variables of mixed type. *Journal of Multivariate Analysis, 178*, 104601. https://doi.org/https://doi.org/10.1016/j.jmva.2020.104601

Ossa Galvis, M. M., Bhakta, R. T., Tarmahomed, A., & Mendez, M. D. (2021). Cyanotic Heart Disease. In *StatPearls*. StatPearls Publishing
Copyright © 2021, StatPearls Publishing LLC.

Paananen, T., Piironen, J., Bürkner, P.-C., & Vehtari, A. (2021). Implicitly adaptive importance sampling. *Statistics and Computing, 31*(2), 16. https://doi.org/10.1007/s11222-020-09982-2

Paleti, R., Bhat, C. R., & Pendyala, R. M. (2013). Integrated Model of Residential Location, Work Location, Vehicle Ownership, and Commute Tour Characteristics. *Transportation Research Record, 2382*(1), 162-172. https://doi.org/10.3141/2382-18

Pedregosa, F. a. V., G. and Gramfort, A. and Michel, V., and Thirion, B. a. G., O. and Blondel, M. and Prettenhofer, P., and Weiss, R. a. D., V. and Vanderplas, J. and Passos, A. and, & Cournapeau, D. a. B., M. and Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *journal of machine learning research, 12*, 2825-2830.

Peterson, R. (2017). *Estimating normalization transformations with bestNormalize*. Retrieved 04/01/2021 from https://github.com/petersonR/bestNormalize

Prudon, P. (2015). Confirmatory Factor Analysis as a Tool in Research Using Questionnaires: A Critique. *Comprehensive Psychology, 4*. https://doi.org/10.2466/03.CP.4.10

Quinn, K. M. (2004). Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses. *Political Analysis, 12*(4), 338-353.

R Core Team. (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Retrieved 05/06/2021 from http://www.R-project.org/

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(4), 805-827. https://doi.org/https://doi.org/10.1111/j.1467-985X.2006.00426.x

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*(2), 167-190. https://doi.org/10.1007/BF02295939

Ramezani, S., Laatikainen, T., Hasanzadeh, K., & Kyttä, M. (2019). Shopping trip mode choice of older adults: an application of activity space and hybrid choice models in understanding the effects of built environment and personal goals. *Transportation*. https://doi.org/10.1007/s11116-019-10065-z

Ranalli, M., & Rocci, R. (2017). Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis, 110*, 87-102. https://doi.org/https://doi.org/10.1016/j.csda.2016.12.016

Reise, S. P., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative Approaches to Addressing Non-Normal Distributions in the Application of IRT Models to Personality Measures. *J Pers Assess, 100*(4), 363-374. https://doi.org/10.1080/00223891.2017.1381969

Reiter & Walsh, P. C. (2017). *Apgar Score for Newborn Health Assessment: What is an Apgar score?* American Baby & Child Law Centers. Retrieved 06/12/2021 from https://www.abclawcenters.com/practice-areas/diagnostic-tests/apgar-score-for-assessment-of-the-newborn/

Rettner, R. (2013). *Big Babies: Are Heavy Newborns Healthy?* Live Science. Retrieved 06/31/2021 from https://www.livescience.com/35771-big-baby-health-risks.html

Revelle, W. (2017). *psych: Procedures for Personality and Psychological Research.* In (Version 1.7.8.) https://CRAN.R-project.org/package=psych

Rex, B. K. (2016). *Principles and Practice of Structural Equation Modeling, Fourth Edition* (Vol. Fourth edition) [Book]. The Guilford Press.

Rezaei Ghahroodi, Z., Aliakbari Saba, R., & Baghfalaki, T. (2019). Gaussian Copula–based Regression Models for the Analysis of Mixed Outcomes: An Application on Household's Utilization of Health Services Data. *Journal of Statistical Theory and Applications*
18(3), 5. https://doi.org/https://doi.org/10.2991/jsta.d.190306.009

Roosa Tikkanen, M. Z. G., Molly FitzGerald, Laurie Zephyrin. (2020). *Maternal Mortality and Maternity Care in the United States Compared to 10 Other Developed Countries*. Retrieved 11/21/2020 from https://www.commonwealthfund.org/publications/issue-briefs/2020/nov/maternal-mortality-maternity-care-us-compared-10-countries

Rosenthal, H., & Voeten, E. (2007). Measuring legal systems. *Journal of Comparative Economics, 35*(4), 711-728. https://doi.org/https://doi.org/10.1016/j.jce.2007.08.001

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *2012, 48*(2), 36. https://doi.org/10.18637/jss.v048.i02

Samani, E., & Ganjali, M. (2010). A Multivariate Latent Variable Model for Mixed - Data from Continuous and Ordinal Responses with Possibility of Missing Responses. *AAM, 5*(2), 1564-1584.

Sanchez, J. P. (2014). Decentralization as a multifaceted concept: a more encompassing index using bayesian statistics. *Revista Española De Ciencia Política, 34*, 9-34.

Santos, R. d. O., Gorgulho, B. M., Castro, M. A. d., Fisberg, R. M., Marchioni, D. M., & Baltar, V. T. (2019). Principal Component Analysis and Factor Analysis: differences and similarities in Nutritional Epidemiology application [Análise de Componentes Principais e Análise Fatorial: diferenças e similaridades na aplicação em Epidemiologia Nutricional]. *Revista Brasileira de Epidemiologia, 22*.

Shi, J.-Q., & Lee, S.-Y. (2000). Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62*(1), 77-87. https://doi.org/https://doi.org/10.1111/1467-9868.00220

Siddiqui, A., Cuttini, M., Wood, R., Velebil, P., Delnord, M., Zile, I., Barros, H., Gissler, M., Hindori-Mohangoo, A. D., Blondel, B., & Zeitlin, J. (2017, Jul). Can the Apgar Score be Used for International Comparisons of Newborn Health? *Paediatr Perinat Epidemiol, 31*(4), 338-345. https://doi.org/10.1111/ppe.12368

Song, X.-Y., & Lee, S.-Y. (2006). A Maximum Likelihood Approach for Multisample Nonlinear Structural Equation Models With Missing Continuous and Dichotomous Data. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(3), 325-351. https://doi.org/10.1207/s15328007sem1303_1

Song, X., Kang, K., Ouyang, M., Jiang, X., & Cai, J. (2018). Bayesian Analysis of Semiparametric Hidden Markov Models With Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(1), 1-20. https://doi.org/10.1080/10705511.2017.1364968

Song, X. Y., Lu, Z. H., Cai, J. H., & Ip, E. H. (2013). A Bayesian modeling approach for generalized semiparametric structural equation models. *Psychometrika, 78*(4), 624-647. https://doi.org/10.1007/s11336-013-9323-7

Sörbom, D. (1989). Model modification. *Psychometrika, 54*(3), 371-384. https://doi.org/10.1007/BF02294623

Stanford Children's Health. (2021). *The Neonatal Intensive Care Unit (NICU)*. Retrieved 1/6/2021 from https://www.stanfordchildrens.org/en/topic/default?id=the-neonatal-intensive-care-unit-nicu-90-P02389

Tabrizi, E., Bahrami Samani, E., & Ganjali, M. (2020). General location multivariate latent variable models for mixed correlated bounded continuous, ordinal, and nominal responses with non-ignorable missing data. *Journal of Applied Statistics*, 1-21. https://doi.org/10.1080/02664763.2020.1745765

Thmasebinejad, Z., & Tabrizi, E. (2015). Sensitivity Analysis in Correlated Bivariate Continuous and Binary Responses. *AAM, 10*(1), 609-619.

Tillmann, J., Uljarevic, M., Crawley, D., Dumas, G., Loth, E., Murphy, D., Buitelaar, J., & Charman, T. (2020). Dissecting the phenotypic heterogeneity in sensory features in autism spectrum disorder: a factor mixture modelling approach. *Mol Autism, 11*(1), 67. https://doi.org/10.1186/s13229-020-00367-w

University of Rochester Medical Center Rochester. (2020). *Newborn Measurements*. Retrieved 05/16/2021 from https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=90&contentid=P02673

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review, 25*(1), 143-154. https://doi.org/10.3758/s13423-016-1015-8

Varriale, R., & Vermunt, J. K. (2012, Mar 30). Multilevel Mixture Factor Models. *Multivariate behavioral research, 47*(2), 247-275. https://doi.org/10.1080/00273171.2012.658337

Vehtari, A., Gelman, A., & Gabry, J. (2016). 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

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

Wall, M. M., Guo, J., & Amemiya, Y. (2012, Mar 30). Mixture Factor Analysis for Approximating a Nonnormally Distributed Continuous Latent Factor With Continuous and Dichotomous Observed Variables. *Multivariate behavioral research, 47*(2), 276-313. https://doi.org/10.1080/00273171.2012.658339

Wang, X., Feng, X., & Song, X. (2020). Joint analysis of semicontinuous data with latent variables. *Computational Statistics & Data Analysis, 151*, 107005. https://doi.org/https://doi.org/10.1016/j.csda.2020.107005

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *journal of machine learning research, 11*(116), 3571-3594.

Wendt, L. P., Wright, A. G. C., Pilkonis, P. A., Nolte, T., Fonagy, P., Montague, P. R., Benecke, C., Krieger, T., & Zimmermann, J. (2019, Nov). The latent structure of interpersonal problems: Validity of dimensional, categorical, and hybrid models. *Journal of abnormal psychology, 128*(8), 823-839. https://doi.org/10.1037/abn0000460

Woods, C. M., & Edwards, M. C. (2011). 6 - Factor Analysis and Related Methods. In C. R. Rao, J. P. Miller, & D. C. Rao (Eds.), *Essential Statistical Methods for Medical Statistics* (pp. 174-201). North-Holland. https://doi.org/https://doi.org/10.1016/B978-0-444-53737-9.50009-8

Woolner, A. M., Ayansina, D., Black, M., & Bhattacharya, S. (2019). The impact of third- or fourth-degree perineal tears on the second pregnancy: A cohort study of 182,445 Scottish women. *PLoS One, 14*(4), e0215180-e0215180. https://doi.org/10.1371/journal.pone.0215180

World Health Organization. (2004a). *ICD-10 : international statistical classification of diseases and related health problems / World Health Organization*. World Health Organization.

World Health Organization. (2004b). *Maternal, newborn, child and adolescent health: What is quality of care and why is it important?* Retrieved 11/21/2020 from https://www.who.int/maternal_child_adolescent/topics/quality-of-care/definition/en/

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods, 51*(1), 409-428. https://doi.org/10.3758/s13428-018-1055-2

Xu, X., & von Davier, M. (2008). FITTING THE STRUCTURED GENERAL DIAGNOSTIC MODEL TO NAEP DATA. *ETS Research Report Series, 2008*(1), i-18. https://doi.org/10.1002/j.2333-8504.2008.tb02113.x

Yang, M., & Dunson, D. B. (2010). Bayesian Semiparametric Structural Equation Models with Latent Variables. *Psychometrika, 75*(4), 675-693. https://doi.org/10.1007/s11336-010-9174-4

Yuan, K.-H. (2016). Meta analytical structural equation modeling: comments on issues with current methods and viable alternatives. *Research Synthesis Methods, 7*(2), 215-231. https://doi.org/10.1002/jrsm.1213

Yuan, K.-H., & Chan, W. (2016). Structural Equation Modeling With Unknown Population Distributions: Ridge Generalized Least Squares. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(2), 163-179. https://doi.org/10.1080/10705511.2015.1077335

Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modelling with correlation matrices for ordinal and continuous data. *The British journal of mathematical and statistical psychology, 64*(Pt 1), 107-133. https://doi.org/10.1348/000711010X497442

Zajic, A. (2019). *Introduction to AIC — Akaike Information Criterion*. Retrieved 05/18 from https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced

Zhao, S., Engelhardt, B. E., Mukherjee, S., & Dunson, D. B. (2018). Fast Moment Estimation for Generalized Latent Dirichlet Models. *Journal of the American Statistical Association, 113*(524), 1528-1540. https://doi.org/10.1080/01621459.2017.1341839

Zhao, Y., & Udell, M. (2020). *Missing Value Imputation for Mixed Data via Gaussian Copula* Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA. https://doi.org/10.1145/3394486.3403106

Zhou, L., Lin, H., Song, X., & Li, Y. I. (2014, Dec). Selection of latent variables for multiple mixed-outcome models. *Scand Stat Theory Appl, 41*(4), 1064-1082. https://doi.org/10.1111/sjos.12084