Examining sources of heterogeneity between studies of mental-health outcomes in children with experience of foster care – a meta-analytical approach

Master thesis in Public Health Sciences (30 ECTS)
Spring 2018
Name: Henrik Karlsson
Supervisors: Lars Brännström
Ylva Almquist

Department of Public Health Sciences
Abstract

Systematic reviews of the effect of foster care on mental-health outcomes have consistently indicated a zero-sum game, which makes it unclear whether the intervention is suitable for children in need of out-of-home placements. This thesis took on a meta-analytical approach to examine sources of heterogeneity between studies evaluating the effect of foster care on adaptive functioning, cognitive functioning, externalizing behavior, internalizing behavior, and total problems behavior. The bulk of studies came from two recently published systematic reviews. The searches were replicated to cover studies published until March 31, 2018. From 2943 studies assessed for eligibility, 240 were selected for the analysis covering 25 085 children. A choice of study-related covariates was abstracted, and potential sources of heterogeneity were hypothesized and tested by means of meta-regression. The findings indicated that both the choice of study design and measurement instrument were significantly associated with the variation in effect sizes. These associations were even stronger in child protection-oriented welfare regimes while insignificant in family service-oriented regimes, which instead showed significant associations for outcome type and publication year. The results imply a need to standardize effect studies of foster care, and to further research on sources of heterogeneity in different child welfare regimes.

Key words
foster care; mental health; sources of heterogeneity; meta-regression
# Table of contents

## Introduction .................................................................................................................. 1

## Aim and research questions ......................................................................................... 3

## Sources of heterogeneity in studies of the effect foster care ........................................ 3

## Methods .......................................................................................................................... 7

## Testing hypotheses ......................................................................................................... 7

## Creating the dataset ....................................................................................................... 7

## Searching of databases ................................................................................................. 7

## Including and excluding studies .................................................................................... 8

## Analyzing the data .......................................................................................................... 9

### Synthetizing results using meta-analysis ................................................................. 9

### Testing for publication bias ....................................................................................... 10

### Identifying and quantifying heterogeneity .............................................................. 10

### Examining sources of heterogeneity using meta-regression .................................... 11

### Study-related covariates ......................................................................................... 12

## Results .......................................................................................................................... 17

## Pooled effect estimates ............................................................................................... 17

## Sources of heterogeneity ............................................................................................ 28

### Welfare regime stratified analyses ........................................................................... 33

## Discussion ..................................................................................................................... 36

## Strengths and weaknesses ........................................................................................... 38

## Conclusions ................................................................................................................... 39

## References .................................................................................................................... 41
Introduction

Child protection authorities place large numbers of young people into foster care all over the world. About 270,000 children entered care in the USA between 1 October 2014, and 30 September 2015 (U.S. Department of Health and Human Services, 2016), and in England about 53,000 children were living with foster families on 31 March 2017 (Ofsted, 2018). About 4 percent of the Swedish youth population experienced foster care between 1992 and 2002 (Franzén et al., 2008), and about 10,000 children entered care in Sweden during 2013 (Socialstyrelsen, 2014).

In the last decades, policymakers have intensified the attempts to transform the child protection practice into an evidence-based practice (Léveillé & Chamberland, 2010), and to increase the use of evidence-based interventions in public health (Brownson et al., 2017). To succeed with this project, practitioners and decision-makers need guidance on which interventions that work the best for various types of foster care children in different circumstances and contexts (Brownson, Fielding & Maylahn, 2009). As foster care children often experience social exclusion and marginalization in adulthood (Stein, 2006), the research community needs to improve the understanding for where and when foster care is a suitable intervention. Foster care is generally the preferred intervention for children that cannot stay at home for various reasons as it resembles a normal family environment (White, 2002).

Researchers have nevertheless questioned the effectiveness of foster care for a long time, despite the popularity among politicians and professional social workers. In the eighties, vulnerable children in the UK got worse living conditions when placed into care, and they experienced more adverse outcomes in young adulthood than children in the majority population (Parker et al., 1991). Similar results have recently been found for former foster care children in the Nordic countries (Kääriäila & Hiilamo, 2017). Later research has also found that former foster care children have poorer outcomes than similar children that were never placed into care, both in the Anglo-Saxon (Doyle 2007, 2008; Warburton et al., 2011), and Nordic countries (Frederiksen 2012; Fuglsang Olsen et al., 2012). In fact, there are no substantial differences in long-term mental-health outcomes between former foster care children in the Nordic, Central European, or Anglo-Saxon countries (Gypen et al., 2017), despite the ability of the welfare state to mitigate social exclusion and promote favorable life outcomes (Esping-Andersen et al. 2012; Ferrarini et al. 2010). Researchers specifying their field of research to mental health-related outcomes have found that foster children more often develop long-term mental health problems involving externalizing behavior (e.g. Clausen et al. 1998; Pottick et
al. 2005) and internalizing behavior (e.g. Lohaus et al., 2017), as well as poorer adaptive functioning (e.g. Webb et al., 2010) and cognitive functioning (e.g. Fry et al., 2017).

Externalizing behavior problems affect people in the social environment, and include aggression, delinquency, and hyperactivity (Achenbach, 1991; Horwitz et al., 2001). It may involve physically or verbally harming someone, or threaten to do so, but also antisocial actions such as theft, burglary, robbery, vandalism, drug abuse, and violence. Hyperactivity refers to two types of problems; the first type refers to an excess of motor activity or restlessness, and the second to attention deficits often experienced in controlled settings such as schools (Liu, 2004). Internalizing behavior problems instead refers to psychological and emotions problems such as anxiety, depression, withdrawal and somatic complaints, which instead primarily harms the person experiencing these adversities (Achenbach, 1991). Adaptive functioning refers to the level of personal independence and social functioning expected from someone given that person’s age according to the dominant cultural proper standards in society (Horwitz et al., 2001). Cognitive functioning refers to intellectual activities and mental processes, such as remembering things, showing attention to something, learning new things, to process various matters, and to use the working memory (González-Ortega et al., 2017). These outcomes have so far been the most commonly used in research on the effect of foster care on mental health-related outcomes (Barber & Delfabbro, 2005; Fanshel & Shinn, 1978; Minty, 1999; Stahmer et al., 2009).

A common approach in public health is to synthesize individual estimates of the effect of an intervention into a summary effect size. The summary effect size will have a better precision as it is based on a larger sample of study participants (Harris et al., 2001; Rychetnik et al., 2004). This approach is called meta-analysis and has been helpful in fields experiencing ambiguous and contradictory findings (Borenstein et al., 2009; Deeks et al., 2008). Recently, Goemans and colleagues (2015, 2016) synthesized decades of research showing that (1) children did not improve their adaptive functioning, externalizing behavior, or internalizing behavior while in care, (2) foster care children had generally lower levels of functioning than children in the majority population, and roughly the same level of functioning as children at risk who stayed at home. In evidence based public health, interventions are developed, implemented and evaluated using scientific methods and principles of reasoning. Data information systems are created to follow-up interventions and document the results. Behavioral science theory and program planning models are applied when appropriate, and local communities are invited to partake in the decision making. The practice seeks the best available scientific
evidence for the administered interventions, feedback from the service providers, and an understanding of the wishes and beliefs held by the service takers. A plan of action is formulated for each service taker (Brownson et al., 2017, 2009).

Interventions are structured programs or activities that aim to improve the health of a population (Brownson et al., 2003), and are best being evaluated in randomized control trials (RCTs) (Harris et al., 2001), a study design that has several advantages such as making counterfactual inference possible by isolating the treatment effect from other factors that may affect the outcome (Sundell, 2008). The best available scientific evidence is the accumulated body of evidence for the effectiveness of an intervention (Brownson et al., 1995). However, a problem with standard meta-analysis is that the method does not sufficiently account for heterogeneity between the studies included in the analysis, which may compromise the validity of the summary effect size (Borenstein et al., 2009; Deeks et al., 2008). So far, the question of heterogeneity has not been discussed concerning studies of mental-health related outcomes among foster care children, despite the extensive amount of available research.

Aim and research questions

By means of meta-analytical tools, this thesis aims to further our understanding regarding potential sources of heterogeneity between studies of mental health outcomes in children with experience of foster care. This is achieved by examining whether and to what extent various hypothesized sources of heterogeneity are linked to variation in effect sizes from 240 studies that in various ways have addressed whether foster care improves outcomes related to adaptive functioning, externalizing behavior, and internalizing behavior. Doing so, this thesis may contribute to improving evidence-based interventions for children who cannot stay at home, and providing guidance to policymakers wanting the best solutions for children in need of care outside of the home in their local area.

Sources of heterogeneity in studies of the effect foster care

This section will briefly outline some common sources of heterogeneity that to a varying degree may explain differences in outcome between individual studies concerning the effectiveness of foster care.

Public health and social care interventions may be viewed as the attempt from society to address certain problems and circumstances that threaten the health of its members, for example, children who grow up in home environments deemed unacceptable and as risking their health and development. As such, these interventions may not only be described as affected
by the constantly changing cultural, political, and social contexts, but as plans of action em-
bodying underlying assumptions and theoretical ideas dominant in a society (e.g. Chisala,
2015; Reisch & Jani, 2012). How societies view and treat foster care children change over
time and context, and so do foster parents’ attitudes, values, and ideas about parenting, as
well as foster care children’s views and understanding of themselves (e.g. Goodyer, 2014;
Whiting, 2000). The history of child welfare systems in different countries have been de-
scribed extensively in the literature, for example, Parton (2014) has authored an interesting
walk-through of early and late changes to the child protection system in England, and
Lundström (1993) has described the major early developments of the Swedish child welfare
system. It is not far-fetched to think that these changes also affect the outcome of the pro-
vided care; hence, individual effect studies using the same study design and procedures could
therefore find different effects depending on time and context.

According to the broad typology suggested by comparative research, child welfare re-
gimes are either child protection or family service oriented (Gilbert, 1997, 2012; Gilbert et
al., 2011). Professional social workers in child protection-oriented regimes tend to adopt a
more legalistic approach, delay interventions, and try to control deviant behavior. They view
child protection services as distinct from other services helping children and families with
needs, and they tend to only remove children from home likely to suffer significant harm
from abuse, defined as violent acts from which the child needs protection. Professionals in
family service-oriented regimes instead tend to regard their work and services as part of the
general public attempts to support families and children. They try to cooperate with families
to improve living conditions and help them preserve social cohesion in hope of not having to
remove children from home. Abuse is regarded as family conflicts and dysfunctions coming
from psychological difficulties and socioeconomic stress, which respond to various forms of
interventions of a psychological character (Gilbert, 2012), such as family treatment (e.g.
Svendsen, 2016). Comparative research has also revealed that these differences in approach
seem to influence which children are removed from home on a systematic level. Child protec-
tion-oriented regimes most often place young children subjected to neglect and abuse, while
family service-oriented regimes most often place teenagers that cannot stay at home because
of emotional and behavioral reasons. Moreover, involuntary placements are more common in
child protection-oriented regimes (Healy et al., 2011). This potential source of heterogeneity
has recently become more important because policymakers in both Nordic and Anglo-Saxon
countries has attempted to strike a new balance between protecting children and providing
services to families in need (Gilbert et al., 2011; Östberg, 2012).
After how long time in care researchers decide to follow-up on the outcomes may also explain some variation in estimated effect sizes between individual studies. On the one hand, it could be argued that positive effects could not take place until the children have handled the loss of the primary caregivers, but on the other hand, positive effects could take place early as it can be assumed that they are provided with better home environments while in care (c.f. Schneider & Phares, 2005). However, when The Swedish Agency for Health Technology Assessment and Assessment of Social Services (2010) synthesized the effect from preventive interventions targeting the same outcomes as in this thesis, they noted that studies with shorter follow-up times were generally reporting larger effect estimates, which makes it difficult to understand whether mental health-related development among children are most effected on the long or short term.

When in childhood the placement takes place may also affect the outcome, and the outcomes might differ between boys and girls. So far, research has revealed that boys and teenagers tend to experience poorer outcomes than girls and pre-teens (e.g. Brännström et al., 2018; Lindquist & Santavirta, 2014; Oosterman et al., 2006), and that children of a higher age when entering care experience higher levels of behavioral problems (McWey et al., 2010).

Attrition is a common concern in RCT-studies as the study participants lost to follow-up might differ systematically from the remaining participants, which potentially could bias the result. For example, it is not far-fetched that more motivated participants stay in a study, which could exaggerate the intervention effect (Dumville et al., 2006).

The proportion of females is important to keep track of as former female foster care children tend to show signs of poorer mental health than their male counterparts (e.g. Coutney et al., 2007).

It is important to choose wisely between the available measurement instruments of mental health-related problems among children and adolescents. In general, it is not enough to choose a psychometrically validated instrument as there might be systematical discrepancies between the theoretical underpinnings of an instrument and the worldview of the study participants. If unaccounted for, this might jeopardize the internal validity of the summary effect (Records et al., 2014), but despite the importance of addressing this issue in order to ensure reliable effect estimates, researchers seldom elaborate on why they chose certain instruments for certain studies. Unintendedly, this could mean that the effect of foster care might be associated with the choice of measurement instrument.
Who responds to the instrument might also affect the measured effect. Some instruments are used with foster parents, other with teachers, and some with the foster children themselves. For smaller children such as toddlers, researchers have to solely rely on adults. Therefore, it cannot be ruled out that the type of respondent is a potential source of heterogeneity.

How researchers design effect studies may also play a role for the estimated effect size. The effect of care may be calculated as the mean difference between baseline and follow-up, or as the mean difference between an intervention group and a comparison group. In this case, the first way will involve only one sample of foster care children, and the second way will involve one group of foster care children that were compared with either similar non-placed children or children in the majority population. In the field of psychotherapy and social work, an intense debate has taken place concerning the method of meta-analysis because of the absence of statistically significant differences in mean effect between bona fide interventions. Several researchers have suggested that psychosocial interventions share the therapeutic alliance between caregivers and caretakers as the active component (Bergmark & Lundström, 2011). Thus, if bona fide interventions are evaluated in randomized trials, the effect size will be significant if participants are compared with baseline or waiting list (receiving no intervention), but insignificant if compared with receivers of another bona fide intervention (Bergmark et al., 2011). The choice of study design is therefore an important source of heterogeneity to examine.

Publishers of scientific journals prefer manuscripts written in English and studies with larger sample sizes able to present statistically significant results (Jüni et al, 2002). Instead of trying to publish insignificant results, researchers tend to keep them in the file drawer (Rosenthal, 1979), which may threaten the validity of the synthetization by exaggerating the size of the summary effect. This phenomenon is called publication bias (Dickersin, 1990).

In light of the above, this thesis has formulated a set of hypotheses for a choice of study-related covariates that were available for extraction from the studies selected for the analysis. The study-related covariates concerns study characteristics, study methodology and design, as well as external factors. In the next section, they will be outlined in detail, and then summarized in Table 3.
Methods

This section outlines the procedures and methods applied in this thesis and goes into detail with the study-related covariates as well as the hypotheses formulated for each one of them.

Testing hypotheses

This thesis applies the 95 percent confidence level in hypothesis testing for statistical significance. This means that the null hypothesis is rejected when the likelihood of obtaining the sample distribution is less than 5 percent, given that the null hypothesis is true (cf. Gigerenzer, 2004). Continuous estimates with confidence intervals not overlapping zero are also considered statistically significant.

Creating the dataset

The first step in creating the dataset was to record the available effect estimates and study-related covariates reported by Goemans and colleagues (2015, 2016) into a spreadsheet using Microsoft Excel (2016). The second step was to prepare the data for analysis using SPSS version 24 (IBM Corp, 2016). The third step was to transfer the data to Stata 15/IC (StataCorp, 2017), and thereafter creating the necessary dummy variables.

Searching of databases

This thesis replicated the searches performed by Goemans and colleagues (2015, 2016) to include studies published after the respective time periods. ERIC, Medline, PsycINFO, and ProQuest Dissertations & Theses were searched. In the last database, the focus was on finding studies with smaller samples and insignificant results, such studies are sometimes found in dissertations and theses (Franco et al., 2014). The replicated search yielded 2939 studies. In addition, 5 other studies from other sources were identified as plausible matches. Thus, a total of 2943 articles were assessed for eligibility. Of these studies, 67 were considered eligible for inclusion, but 51 of them lacked the necessary information for an effect size to be calculated. The authors of these studies were emailed with a request of additional information, which resulted in two answers providing the necessary information. Subsequently, 18 studies from the replicated search were included in the meta-analysis for a final sample of 240 studies (k = 240) that covered 25 085 foster care children (n = 25 085), see Table 1 and Figure 1.
Including and excluding studies

This thesis used the same inclusion and exclusion criteria as Goemans and colleagues (2015, 2016). First, eligible studies had to include children aged between 0 and 18 years placed in foster care outside of the home. Studies were included regardless if comparison groups had been used or not. If longitudinal studies had measured the effect at several time points, the effect size was calculated using the two time points the farthest away. Studies that had used norm scores to represent the majority population were excluded.

Second, studies had to have reported on the following mental health-related outcomes; adaptive functioning, cognitive functioning, externalizing behavior, internalizing behavior, and total problems behavior. Studies that reported on the proportion of children scoring clinically significant results for various types of diagnosis were excluded as straightforward conclusions cannot be drawn from such tests (Hudziak et al., 2004).

Third, if the effect size had not been reported as Hedges’ $g$, enough information had to have been reported for it to be calculated in Lenhard and Lenhard’s (2016) online effect size calculator. Hedges’ $g$ uses pooled weighted standard deviations instead of pooled standard deviations, which makes it more reliable than Cohen’s $d$ for studies smaller than 20 participants (Durlak, 2009; Ellis, 2010).

Fourth, studies written in other languages than English were not excluded in beforehand, but if they were not written in Swedish, Danish, or Norwegian, they had to provide an English summary with enough information for the eligibility assessment.

Table 1

| Search strategy for ERIC, Medline, PsycINFO, and ProQuest Dissertations & Theses |
| Search string | Search string |
| ("foster children" OR "foster care") AND (longitudinal OR "reported measures" OR "pretests posttests") | ("foster child**" OR "foster care") AND (internalizing OR externalizing OR behavio* OR SDQ OR VABS OR CBC* OR development* OR disorder* OR cognitive OR IQ OR intelligence OR intellect* OR mental problem* OR "mental health" OR psychological problem* OR compar* OR psychopatho* OR Vineland OR adaptive) |
| Time span | Time span |
| January 1, 2014 to March 31, 2018 | May 1, 2014, to March 31, 2018 |
Analyzing the data

Analyzes and calculations were performed using the {\texttt{mais}} package (Sterne, 2009) available for Stata 15/IC (StataCorp, 2017). One extreme value was dropped for the meta-analyses, leaving a total of 239 observations \((k = 239)\) in the analytical sample.

Synthesizing results using meta-analysis

Two meta-analytical methods were applied. First, the {\texttt{metan}} command was used to perform a random-effects meta-analysis to synthetize the effect estimates. As true heterogeneity between the effect sizes was assumed, random effects-models were applied to allow the true effect size to vary between the studies. Thus, the random-effects model rather calculates a summary effect size in terms of a synthesized mean of estimated effect sizes (Borenstein et al., 2009; Huedo-Medina et al., 2006).

Second, the {\texttt{metacum}} command was used to perform random-effects cumulative meta-analyses, which tracks the accumulation of evidence of an intervention as research progress (Lau et al., 1992). This is used to illustrate the zero-sum game of evaluating the effect of foster care on mental-health outcomes.
Testing for publication bias

The bias assessment plot, also called the funnel plot, was used to test the included studies for publication bias. On the x-axis, each study’s effect size estimate is plotted, and on the y-axis, each estimate’s standard deviation is plotted. The pseudo 95 confidence limits are visualized in the graph, forming an inverted funnel with the x-axis. As the sampling error is random, publication bias is indicated by either one of the spaces to the left or the right of the mean being blank (Borenstein et al., 2009). In addition, the linear association between the effect size estimates and their standard errors was assessed and visualized by Egger’s regression line (Egger et al., 1997).

Identifying and quantifying heterogeneity

The true effect size refers to the effect size in the population from which the sample was drawn, and the observed effect size refers to the effect size observed in the sample. The terms variation and dispersion describe differences between observed values, and heterogeneity describes differences between true effect sizes. A choice of effect studies often shows a pattern of observed effect estimates, rather than a straight line of estimates of a similar size and precision. Identifying and quantifying heterogeneity in true effects is about making sense of such patterns. How the effect estimates vary may have important implications for how to interpret the effectiveness of an intervention. Borenstein and colleagues (2011) suggest using the following questions to get a picture of the heterogeneity:

- Is there evidence of heterogeneity in true effect sizes?
- What is the variance of the true effects?
- What are the substantive implications of this heterogeneity?
- What proportion of the observed dispersion is real?

The observed heterogeneity between estimated effect sizes includes both true variance and random error, which means that the true variance must be isolated. This is, however, not an uncomplicated process. Even if the heterogeneity in true effects was zero, the observed effects would not be identical to each other but rather fall within some range of the true effect due to the random error (Borenstein et al., 2011). So, if effect sizes truly vary between studies, the observed effects would thus vary both due to the random error, and the true heterogeneity of the effect sizes. The first step of isolating the true heterogeneity is to compute the weighted sum of squares, called the $Q$-statistic. If all studies in a sample are assumed to share the same effect size, and all variance is attributable to the random error, the expected value of
the $Q$-statistic is simply the degrees of freedom. The chi square test of the distribution of expected and actual weighted sum of squares determines whether true heterogeneity exists between the included studies (a.a).

In the next step, the variance of effect sizes in the population from which the sample was drawn ($\tau^2$) is estimated from the variance of observed effect ($T^2$). If the deviance around the mean of the true effect sizes is assumed to be normally distributed, the standard deviation of the observed effect ($T$) may be used to get a sense of the distribution of the true effect sizes around the mean. This measure makes it possible to reflect upon how many effect sizes that fall within a certain range. Both $T$ and $T^2$ are absolute measures that quantify deviation using the same scale as the effect sizes (Borenstein et al., 2011).

The $I^2$-statistic is in contrast a relative measure used to quantify the proportion of the observed variance that reflect real differences in effect sizes (Huedo-Medina et al., 2006). To be more specific, it is the ratio of excess dispersion to total dispersion, which reflects the extent to which the confidence intervals of the observed effect sizes overlaps. The $I^2$-statistic ranges from 0 percent to 100 percent (Borenstein et al., 2011). If the proportion is large, meta-regression is a suitable approach for trying to explain the variance. Higgins and Thompson (2002) have proposed a classification of $I^2$ values; percentages of around 25%, 50%, and 75% corresponds to low, medium, and high variance, respectively.

**Examining sources of heterogeneity using meta-regression**

As a first step, simple meta-regression was conducted for each study-related covariate to test the specific hypothesis of association concerning that covariate, or with other words, to test for potential effect moderators. The procedure is similar as in simple linear regression, but the difference is that the dependent variable is the effect estimate of each study, and the independent variables relate to methodological and contextual differences between the included studies (Borenstein et al., 2011; Thompson & Higgins, 2001), and not differences between study participants as such data is seldom available (Higgins & Thompson, 2004).

As a second step, a multivariable random-effects meta-regression was performed including the effect moderators found in step one. This kept the effect moderators from being over-adjusted. There is no need to control for covariates not shown to be effect moderators.

In cases were studies reported on several outcomes, two or more outcomes were subsequently recorded in the dataset, meaning that the same sample was sometimes used more than once. The literature generally recommends researchers not to add the same sample twice in a meta-analysis, but this is generally not a problem when conducting meta-regression as it was
designed to handle several outcome measures from the same study (Harbord & Higgins, 2008).

**Study-related covariates**

This section outlines the study-related covariates that in advance of the study was considered potential sources of heterogeneity. The operationalization of each covariate is explained along with the formulated hypothesis.

Longitudinal studies calculate attrition as the percentage of non-response at the last follow-up. In non-longitudinal studies, that is, studies without follow-up, the attrition was set to zero. It could be that children experiencing care more positively than others more often stay in the study until its conclusion. Assuming this was true, the rate of the attrition would positively predict the effect size, or to put it differently, longitudinal studies were hypothesized as significantly associated with larger effect sizes as children experiencing the care negatively had been sorted out from the samples at the time of follow-up.

Research reviewers has found the length of follow-up periods in RCTs important for the effect of both psychotherapeutic (e.g. Shean, 2014) and medical interventions (e.g. Llewellyn-Bennett et al., 2016). Goemans and collegues (2015) saw in their meta-analysis that studies with larger samples and longer follow-up periods tended to report smaller effects. From this perspective, in contrast to the hypothesis concerning attrition, it is hypothesized that foster care has poorer long-term than short-term effects. The follow-up time were operationalized as the number of months.

The age of the foster care children has been linked to higher levels of behaviour problems (McWey et al., 2010), and children placed in foster care in their teens have on average poorer outcomes than children placed earlier in childhood (Brännström et al., 2018). This motivated the inclusion of the mean age of the study participants with the hypothesis of a significant association between higher age and poorer outcomes.

The percent of females in the study sample was included as a covariate. The hypothesis given the indications of earlier research (e.g. Courtney et al., 2007), was that a larger percentage of female study participants would be on average associated with smaller effect sizes.

The choice of comparison group was operationalized as follows. The study results had either been compared with a group of similar non-placed children or with the majority population. Some studies were designed without a comparison group to instead calculate the mean difference in outcome between two or more time points. The hypothesis was that there were
no average differences effect sizes between studies using either design as the choice of comparison group was not expected to influence the care provided and the outcome of it.

Sample size was included as a covariate as the $p$-value is related to the sample size. In general, it is easier to obtain statistically significant results with larger sample sizes, which motivated the hypothesis of larger samples being associated with larger effect sizes.

Goemans and colleagues (2015, 2016) could not find significantly better results for any of the mental-health outcomes studied. Subsequently, the hypothesis was that there would be no difference in effect size depending on the type of outcome measured.

In the included studies, a wide arrange of measurement instruments, and combinations of instruments, had been used. They were collapsed into the following categories; Child Behavior Checklist (CBCL); CBCL combined with another instrument; Teachers’ Report Form (TRF); and Other instruments, a category that in turn included 19 instruments, see Table 2. No association between the measurement instrument and the effect of care was hypothesized.

The respondents of the instruments also varied. In some studies, the foster care children themselves responded to the questions, while other studies relied on foster parents or teachers to provide assessments. Some studies also used a combination of respondents. Information on the respondent was abstracted from the individual effect studies, however, the covariate was shown to be heavily correlated with the measurement instrument used as many instruments are designed for usage with a certain type of respondent ($r = 0.467; k = 239; p = .000$). For this reason, this covariate was excluded from the analysis.

The type of publication was divided into peer-reviewed and non-peer reviewed. The hypothesis was that effect estimates published in peer-reviewed journals were significantly higher than those published in non-peer reviewed journals, book chapters, and dissertations.

According to the broad typology suggested by Gilbert (1997), the studies were classified as having been performed either in child protection-oriented child welfare regimes or family services-oriented regimes. Countries classified as child protection-oriented were Australia, Canada, England, Northern Ireland, Scotland, United Kingdom, and USA. Countries classified as family service-oriented were Belgium, Finland, Germany, Iraq, the Netherlands, Norway, Serbia, South Korea, Spain, and Turkey. The differences in cultural, political and social contexts motivate the inclusion of this study-related covariate; nevertheless, the type of child welfare regime was not expected to be associated with the intervention effect because Gypen and colleagues (2017) saw no differences in the level of adversities facing former foster care children in neither of the regimes. As a complementary analysis, this covariate was used to
stratify the analysis to examine whether the strength and pattern of potential sources heterogeneity were the same in both type of child welfare regimes.

The publication year of the study was included because the practice of foster care and the outcomes from the care have most likely been improved since the 70s when the oldest study of this sample was published. In fact, this thesis covers a time span of 60 years, which emphasizes the need for this covariate.

Table 3 outlines the complete list of all study-related covariates along with their name, respective operationalization, and formulated hypothesis.
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Short description</th>
<th>n(^a)</th>
<th>Collapsed category</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASC-2</td>
<td>Behavior Assessment System for Children, Second Edition (BASC-2) is an assessment designed for use in evaluating emotional and behavior disorders in children and adolescents with cognitive, emotional, or learning disabilities (Reynolds &amp; Kamphaus, 2004).</td>
<td>5</td>
<td>Other</td>
</tr>
<tr>
<td>CASAFS</td>
<td>Child and Adolescent Social and Adaptive Functioning Scale (CASAFS) is a self-report measure designed to examine the social functioning of children and adolescents in the areas of school performance, peer relationships, family relationships, and home duties/self-care (Price et al., 2002).</td>
<td>5</td>
<td>Other</td>
</tr>
<tr>
<td>CBCL</td>
<td>Child Behavior Checklist (CBCL) is a report form for caregivers helping to find problem behavior in children (Aschenbach, 1991).</td>
<td>149</td>
<td>CBCL</td>
</tr>
<tr>
<td>CDI</td>
<td>Children's Depression Inventory (CDI) is a self-report measure for cognitive, affective, behavioral problem among children showing signs of depression (Saylor et al., 1984).</td>
<td>22</td>
<td>Other(^b)</td>
</tr>
<tr>
<td>ITSEA</td>
<td>The Infant–Toddler Social and Emotional Assessment (ITSEA) screens for social-emotional and competency developmental delays in infants and toddlers (Carter et al., 2003).</td>
<td>9</td>
<td>Other</td>
</tr>
<tr>
<td>TRF</td>
<td>Teacher’s Report Form (TRF) is a measurement tool for evaluating behavior problems in children displayed as school, for example, delinquency and aggression (Aschenbach, 1991).</td>
<td>19</td>
<td>TRF / CBCL + Other</td>
</tr>
<tr>
<td>VABS</td>
<td>The Vineland Adaptive Behavior Scales (VABS) is a survey interview with foster parents and/or teachers to measure the child’s level of personal and social skills needed for everyday living (Sparrow, 2011).</td>
<td>12</td>
<td>Other / CBCL +</td>
</tr>
<tr>
<td>WISC</td>
<td>Wechsler Intelligence Scale for Children (WISC) tests the intellectual ability of children aged six to sixteen years. Areas such as verbal function, perceptual function, working memory, and speed are covered. Several further developments of this instrument have been released, some of which have been used in studies included in this thesis, i.e. WISC-R, WISC-III, WPPSI, and WPPSI-R.</td>
<td>18</td>
<td>Other(^c)</td>
</tr>
<tr>
<td>YSR</td>
<td>Youth Self Report (YSR) is an instrument for children aged 11 to 18 years of age to self-report social competencies and behavioral problems in a standardized manner (Aschenbach, 1991).</td>
<td>17</td>
<td>Other / CBCL +</td>
</tr>
</tbody>
</table>

\(^a\) Instruments used in less than five studies were: Beck Anxiety Inventory (BAI), Bayley Scales of Infant and Toddler Development (BSID-II), Center for Epidemiological Studies—Depression Scale, Early Child Behavior Inventory (ECBI), Mullen Scales of Early Learning (MSEL), Multidimensional Anxiety Scale for Children (MASC), Performance Development Review (PDR), Pediatric Quality of Life (PedsQL), Rutter Behavior Scales, Self Perception Profile for Children (SPPC), and Strengths and Difficulties Questionnaire (SDQ).

\(^b\) Number of studies which had used the instrument. In several cases it may have been used in combination with one or more other instruments.

\(^c\) This instrument was used too often in combination with another instrument to make up for an own category.

\(^d\) Besides from c, there were too many versions of this instrument for it to make up a clearly distinguishable category.
Table 3
List of study-related covariates along with their respective operationalization and formulated hypothesis

<table>
<thead>
<tr>
<th>Variable</th>
<th>Operationalization</th>
<th>Categories</th>
<th>Hypothesized assoc. w/ ES&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attrition</td>
<td>The size of the attrition at the last follow-up (T2)</td>
<td>Continuous</td>
<td>Higher attrition positively associated</td>
</tr>
<tr>
<td>Length of follow-up</td>
<td>The length of follow-up in months</td>
<td>Continuous</td>
<td>Longer follow-up negatively associated</td>
</tr>
<tr>
<td>Mean age&lt;sup&gt;c&lt;/sup&gt;</td>
<td>The mean age of the children at the start of the study (T1)</td>
<td>Continuous</td>
<td>Higher age negatively associated</td>
</tr>
<tr>
<td>Percent females&lt;sup&gt;d&lt;/sup&gt;</td>
<td>The proportion of females in the study sample</td>
<td>Continuous</td>
<td>Higher percent negatively associated</td>
</tr>
<tr>
<td>Sample size</td>
<td>The size of the sample at the beginning of the study interval (T1)</td>
<td>Continuous</td>
<td>Larger samples positively associated</td>
</tr>
<tr>
<td><strong>Study methodology and design</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison group</td>
<td>What type of comparison group, if any, that had been used.</td>
<td>No comparison group, children at risk staying at home, children in the majority population</td>
<td>No association</td>
</tr>
<tr>
<td>Outcome type</td>
<td>The type of health-related outcome</td>
<td>Adaptive functioning, cognitive, functioning, externalizing behavior, internalizing behavior, total problem behavior</td>
<td>No association</td>
</tr>
<tr>
<td>Measurement instrument</td>
<td>The measurement tool used to collect test scores</td>
<td>CBCL&lt;sup&gt;b&lt;/sup&gt;, CBCL + Other, TRF&lt;sup&gt;c&lt;/sup&gt;, Other</td>
<td>No association</td>
</tr>
<tr>
<td><strong>External factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Welfare regime</td>
<td>The type of welfare regime in the country</td>
<td>Child protection or family service oriented</td>
<td>No association</td>
</tr>
<tr>
<td>Publication type</td>
<td>The type of publication</td>
<td>Peer-reviewed or not peer-reviewed</td>
<td>Peer-reviewed positively associated</td>
</tr>
<tr>
<td>Year</td>
<td>The publication year of the study</td>
<td>Continuous</td>
<td>Newer studies positively associated</td>
</tr>
</tbody>
</table>

<sup>a</sup> ES = Effect size; <sup>b</sup> CBCL = Child Behavior Check List; <sup>c</sup> TRF = Teacher’s Report Form
Results

Pooled effect estimates

Table 4 outlines the results from the standard meta-analysis, which are graphically visualized in Figures 2 to 6. The results indicate a statistically significant effect from foster care on adaptive functioning \((g = -0.160, 95\% CI: -0.267 to -0.052)\), but the clinical relevance is limited as the effect size corresponds to a 90 percent overlap (Becker, 2000). The effect estimates for cognitive functioning, externalizing behavior, internalizing behavior, and total problems behavior is insignificant at the 95 percent confidence level. In Figures 7 to 11, the results from the cumulative meta-analysis are graphically visualized. For all outcome types, the zero-sum game gets even clearer as more precise estimates are synthetized.

The \(Q\)-statistic and the corresponding chi square test indicate true heterogeneity between studies of all outcome types \((p < .00)\). The variance between the true effect sizes is the lowest for adaptive functioning \((\tau^2 = 0.075)\), and the highest for cognitive functioning \((\tau^2 = 0.520)\). Considerable dispersion is indicated for most outcome types by the spread of the true effects around the mean. The lowest dispersion is indicated for adaptive functioning \((\tau = 0.273)\), and the highest for cognitive functioning \((\tau = 0.721)\). This means that some proportion of the effect of foster care on cognitive functioning may be both in the trivial and high range of effect (c.f. Borenstein et al., 2011). A similar situation seems to be the case for externalizing behavior \((\tau = 0.643)\), internalizing behavior \((\tau = 0.466)\), and total problem behavior \((\tau = 0.544)\), while adaptive functioning has less of a wide-ranging distribution of effect sizes \((\tau = 0.273)\). The \(I^2\)-statistic exceeds 99 percent for all meta-analyses, which indicate that a high proportion of the observed variance reflect real differences in effect sizes. This motivates further investigation of sources of heterogeneity between the synthetized studies.

Table 4

<table>
<thead>
<tr>
<th>Outcome</th>
<th>(k)</th>
<th>Pooled estimate(^a)</th>
<th>95 % CI</th>
<th>(Q)</th>
<th>(p_0)</th>
<th>(df)</th>
<th>(\tau^2 (\tau))</th>
<th>(I^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive functioning</td>
<td>28</td>
<td>-0.160</td>
<td>-0.267 to -0.052(^*)</td>
<td>0.00013</td>
<td>.000</td>
<td>27</td>
<td>0.075 (0.273)</td>
<td>99.8%</td>
</tr>
<tr>
<td>Cognitive functioning</td>
<td>10</td>
<td>-0.265</td>
<td>-0.728 to 0.197</td>
<td>3046.021</td>
<td>.000</td>
<td>9</td>
<td>0.520 (0.721)</td>
<td>99.7%</td>
</tr>
<tr>
<td>Externalizing behavior</td>
<td>66</td>
<td>0.128</td>
<td>-0.029 to 0.285</td>
<td>0.00071</td>
<td>.000</td>
<td>65</td>
<td>0.414 (0.643)</td>
<td>99.9%</td>
</tr>
<tr>
<td>Internalizing behavior</td>
<td>65</td>
<td>-0.014</td>
<td>-0.247 to 0.805</td>
<td>0.00037</td>
<td>.000</td>
<td>64</td>
<td>0.217 (0.466)</td>
<td>99.8%</td>
</tr>
<tr>
<td>Total problem behavior</td>
<td>70</td>
<td>0.066</td>
<td>-0.064 to 0.196</td>
<td>0.00088</td>
<td>.000</td>
<td>69</td>
<td>0.296 (0.544)</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

\(^a\) The pooled estimate is the summary effect size calculated as Hedge’s \(g\), i.e. standardized mean difference
Figure 2

Forest plot over the effect of foster care on adaptive functioning, \( k = 28 \)

ES = Effect size calculated as Hedges’ \( g \).
Figure 3

Forest plot over the effect of foster care on cognitive functioning, $k = 10$

ES = Effect size calculated as Hedges’ $g$. 

<table>
<thead>
<tr>
<th>Authors of study</th>
<th>ES (50% CI)</th>
<th>% Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jacobsen et al. (2013)</td>
<td>-0.97 (-1.91, -0.03)</td>
<td>10.70</td>
</tr>
<tr>
<td>Johnson et al. (2010)</td>
<td>-1.56 (-1.61, -1.51)</td>
<td>10.69</td>
</tr>
<tr>
<td>Pears &amp; Fisher (2005)</td>
<td>-0.71 (-0.78, -0.64)</td>
<td>10.68</td>
</tr>
<tr>
<td>Tizard &amp; Hodges (1978)</td>
<td>-1.15 (-2.50, 0.20)</td>
<td>5.59</td>
</tr>
<tr>
<td>Leifer &amp; Shapiro (1995)</td>
<td>0.00 (0.18, 0.18)</td>
<td>10.54</td>
</tr>
<tr>
<td>Maroufi (2003)</td>
<td>0.11 (0.05, 0.17)</td>
<td>10.68</td>
</tr>
<tr>
<td>Min et al. (2014)</td>
<td>0.10 (0.10, 0.01)</td>
<td>10.66</td>
</tr>
<tr>
<td>Sato et al. (2006)</td>
<td>1.41 (0.90, 1.92)</td>
<td>9.48</td>
</tr>
<tr>
<td>Victor et al. (2008)</td>
<td>0.22 (0.25, 0.39)</td>
<td>10.68</td>
</tr>
<tr>
<td>Vidl et al. (1988)</td>
<td>0.23 (0.01, 0.40)</td>
<td>10.30</td>
</tr>
<tr>
<td>Overall ($I^2$ = 99.7%, $p = 0.000$)</td>
<td>-0.27 (-0.73, 0.20)</td>
<td>100.00</td>
</tr>
</tbody>
</table>

NOTE: Weights are from random effects analysis.
Figure 4

Forest plot over the effect of foster care on externalizing behavior, $k = 66$

ES = Effect size calculated as Hedges’ $g$. 
Figure 5
Forest plot over the effect of foster care on internalizing behavior, \( k = 65 \)

ES = Effect size calculated as Hedges’ \( g \).
Figure 6

Forest plot over the effect of foster care on total problem behavior, \( k = 70 \)

ES = Effect size calculated as Hedges’ \( g \).
Figure 7

Cumulative forest plot over the effect of foster care on adaptive functioning

ES=Effect size calculated as Hedges’ g.
Figure 8

Cumulative forest plot over the effect of foster care on cognitive functioning

ES=Effect size calculated as Hedges’ g; GP=General population; Remaining home
Figure 9
Cumulative forest plot over the effect of foster care on externalizing behavior

ES=Effect size calculated as Hedges’ g; GP=General population; Remaining home
Figure 10

Cumulative forest plot over the effect of foster care on internalizing behavior

ES=Effect size calculated as Hedges’ g; GP=General population; Remaining home
Figure 11

Cumulative forest plot over the effect of foster care on total problem behavior

ES=Effect size calculated as Hedges’ $g$, GP=General population; Remaining home
The funnel plot in Figure 12 does not show asymmetry for the naked eye, and the Egger’s test indicate a non-significant relationship between the effect sizes and the precision of the studies ($B = -1.447$, 95 % CI: -6.559 to 4.320; $p = 0.685$). This indicates that publication bias is not present among the included studies.

![Funnel plot with Egger's regression line in orange](image)

**Figure 12**
*Funnel plot with Egger’s regression line in orange*

SE = Standard error

**Sources of heterogeneity**

Table 5 outlines descriptive statistics of the study-related covariates included in the meta-regression. The average effect of foster care on the mental health-related outcomes was $g = 0.027$ ($SD = 0.493$), a small and insignificant effect size (c.f. Becker, 2000).

The mean follow-up time was 12.76 months, and the mean sample size was 104.73 study participants. The mean age of the foster care children was 8.8 years, and about 49 percent females on average. The mean attrition at follow-up was 15.38 percent. Among outcome types, total behavior problem was the most common (29.29 %), followed by externalizing behavior (27.20 %), and internalizing behavior (27.62 %). The most commonly used measurement tool
was CBCL (47.28 %). Most studies had no comparison group (64.43 %). A majority of studies were published in peer-reviewed journals (70.71 %), and most studies had been conducted in child protection-oriented child welfare regimes (63.60 %).

Table 6 outlines the results from the meta-regression analysis performed in step one (crude estimates) and step two (adjusted estimates).

The attrition rate is not associated with the effect size. This result rejects the hypothesis of longitudinal studies as significantly associated with better outcomes, meaning that children who had experienced the care negatively were not sorted out at follow-up.

Neither the covariate for sample size, nor the covariate for the length of follow-up is significant. This rejects the hypotheses of studies having larger sample sizes and longer follow-ups reporting on average smaller effect sizes.

The mean age of the foster care children is not statistically significant, which rejects the hypothesis of a decreasing effect with increasing age.

The percentage of females in the study sample was not significantly associated with smaller effect sizes, which rejects the hypothesis of studies with larger percentages of female study participants reporting on average smaller effect sizes.

Studies that had used children in the majority population as comparison group report on average 0.266 g (95% CI: 0.063 to 0.389) higher effect size than studies having no comparison group. This rejects the hypothesis of no difference between the study designs.

Effect estimates for externalizing behavior are on average 0.138 g (95% CI: 0.003 to 0.274) higher than effect estimates for adaptive functioning, cognitive functioning, internalizing behavior, and total problem behavior problem. But the association is not significant after being adjusted for the predictive validity of the other effect moderators, thus, the hypothesis of similar effect sizes across outcome types is confirmed.

Studies that had used TRF report on average 0.344 g (95% CI: 0.146 to 0.543) higher effect than studies that did not use TRF. Studies that had used another instrument than CBCL, CBCL combined with another instrument, or TRF report on average 0.192 g (95% CI: -0.332 to -0.052) lower effect. But these associations were not significant after being adjusted for the other effect moderators, thus, the hypothesis of no association between the measurement instrument and the size of the effect size is also confirmed.

There was no indication of publication bias, and the covariate for publication type showed an insignificant association with the effect sizes, which rejects the hypothesis of generally larger effect sizes published by peer-reviewed journals compared to non-peer reviewed publications such as book chapters and dissertations.
Studies conducted in child protection-oriented child welfare regimes did not report generally better outcomes, which confirmed the hypothesis of no general differences in measured effect depending on the regime.

Finally, the publication year is not statistically significant, which rejects the hypothesis of later studies reporting on average better results.

In the adjusted model, the between-studies variance is $\tau^2 = 0.194$, and percentage of residual variation due to heterogeneity is $I^2 = 99.83$ percent, which indicate a high heterogeneity among the 239 analyzed studies. The proportion of between-study variance that the covariates explain is about 10 percent ($I^2_{\text{adjusted}} = 10.08 \%$).
# Table 5

**Descriptive statistics, k = 239**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean/Proportion (SD)</th>
<th>Minimum – Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of foster care (dependent)</td>
<td>0.027 (0.493)</td>
<td>-1.56 – 1.41</td>
</tr>
<tr>
<td><strong>Study characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attrition</td>
<td>15.38 (22.657)</td>
<td>0 – 85.42</td>
</tr>
<tr>
<td>Length of follow-up in months</td>
<td>12.76 (16.589)</td>
<td>0 – 63.36</td>
</tr>
<tr>
<td>Mean age(^a)</td>
<td>8.80 (3.368)</td>
<td>0.78 – 17.55</td>
</tr>
<tr>
<td>Percent females(^b)</td>
<td>49.34 (12.249)</td>
<td>0 – 100</td>
</tr>
<tr>
<td>Sample size</td>
<td>104.73 (123.186)</td>
<td>3 – 758</td>
</tr>
<tr>
<td><strong>Study methodology and design</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No comparison group</td>
<td>64.43 (47.971)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Children staying at home</td>
<td>13.81 (34.570)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Children in general population</td>
<td>21.76 (41.346)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Outcome type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptive functioning</td>
<td>11.72 (32.280)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Cognitive functioning</td>
<td>4.18 (20.065)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Internalizing behavior</td>
<td>27.20 (44.591)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Externalizing behavior</td>
<td>27.62 (44.803)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Total behavior problems</td>
<td>29.29 (45.604)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Measurement instrument</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBCL</td>
<td>47.28 (50.031)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>CBCL + Other</td>
<td>17.57 (38.139)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>TRF</td>
<td>10.04 (30.119)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Other</td>
<td>25.10 (43.524)</td>
<td>0 – 1</td>
</tr>
<tr>
<td><strong>External factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publication type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peer-reviewed</td>
<td>70.71 (45.604)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Non-peer-reviewed</td>
<td>29.29 (45.604)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Welfare regime</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child protection oriented</td>
<td>63.60 (48.216)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Family service oriented</td>
<td>36.40 (48.216)</td>
<td>0 – 1</td>
</tr>
<tr>
<td>Year</td>
<td>2006.46 (8.408)</td>
<td>1978 – 2018</td>
</tr>
</tbody>
</table>

\(^a\) \(k = 211\); \(^b\) \(k = 232\); \(^c\) CBCL = Child Behavior Check List; \(^d\) TRF = Teacher’s Report Form
Table 6
The association between study-related covariates and effect estimates, results from random-effects meta-regression analysis, $k = 239$

<table>
<thead>
<tr>
<th>Variable</th>
<th>Crude (95% CI)</th>
<th>Adjusted (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attrition</td>
<td>0.000 (-0.003 to 0.003)</td>
<td>-</td>
</tr>
<tr>
<td>Length of follow-up in months</td>
<td>-0.001 (-0.004 to 0.003)</td>
<td>-</td>
</tr>
<tr>
<td>Mean age$^a$</td>
<td>0.012 (-0.007 to 0.031)</td>
<td>-</td>
</tr>
<tr>
<td>Percent females$^b$</td>
<td>0.001 (-0.004 to 0.006)</td>
<td>-</td>
</tr>
<tr>
<td>Sample size</td>
<td>-0.000 (-0.001 to 0.000)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Study methodology and design</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No comparison group</td>
<td>-0.116 (-0.243 to 0.012)</td>
<td>(reference)</td>
</tr>
<tr>
<td>Children staying at home</td>
<td>0.097 (-0.274 to 0.080)</td>
<td>0.073 (-0.133 to 0.279)</td>
</tr>
<tr>
<td>Children in majority population</td>
<td>0.225 (0.079 to 0.371)$^*$</td>
<td>0.226 (0.063 to 0.389)$^*$</td>
</tr>
<tr>
<td><strong>Outcome type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptive functioning</td>
<td>-0.181 (-0.371 to 0.009)</td>
<td>(reference)</td>
</tr>
<tr>
<td>Cognitive functioning</td>
<td>-0.304 (-0.618 to 0.010)</td>
<td>-0.116 (-0.323 to 0.091)</td>
</tr>
<tr>
<td>Internalizing behavior</td>
<td>-0.053 (-0.190 to 0.084)</td>
<td>-0.229 (-0.575 to 0.117)</td>
</tr>
<tr>
<td>Externalizing behavior</td>
<td>0.138 (0.003 to 0.274)$^*$</td>
<td>0.137 (-0.019 to 0.294)</td>
</tr>
<tr>
<td>Total behavior problems</td>
<td>0.061 (-0.073 to 0.196)</td>
<td>0.073 (-0.084 to 0.229)</td>
</tr>
<tr>
<td><strong>Measurement instrument</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBCL$^a$</td>
<td>0.009 (-0.114 to 0.131)</td>
<td>(reference)</td>
</tr>
<tr>
<td>CBCL + Other</td>
<td>0.016 (-0.143 to 0.175)</td>
<td>-0.044 (-0.230 to 0.141)</td>
</tr>
<tr>
<td>TRF$^b$</td>
<td>0.344 (0.146 to 0.543)$^*$</td>
<td>0.207 (-0.008 to 0.423)</td>
</tr>
<tr>
<td>Other</td>
<td>-0.192 (-0.332 to -0.052)$^*$</td>
<td>-0.142 (-0.296 to 0.013)</td>
</tr>
<tr>
<td><strong>External factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peer-reviewed (ref = non-peer-reviewed)</td>
<td>-0.036 (-0.172 to 0.099)</td>
<td>-</td>
</tr>
<tr>
<td>Child protection oriented (ref = family service oriented)</td>
<td>-0.043 (-0.170 to 0.083)</td>
<td>-</td>
</tr>
<tr>
<td>Year</td>
<td>0.003 (-0.004 to 0.010)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Heterogeneity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent heterogeneity explained</td>
<td>$\tau^2 = 0.194$, $I^2 = 99.83%$,</td>
<td></td>
</tr>
<tr>
<td>$F_{\text{adjusted}} = 10.08%$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$k_{\text{crude}} = 211$; $k_{\text{crude}} = 232$; $^a$ CBCL = Child Behavior Check List; $^b$ TRF = Teacher’s Report Form; $^* p<.05$
Welfare regime stratified analyses

In addition, the meta-regression analysis has been stratified to examine if the strength and pattern of effect moderators were the same in both child protection and family services-oriented child welfare regimes. Table 7 outlines the result.

In child protection-oriented regimes, the mean age of the foster care children is significantly associated with effect of care. For every year increase in mean age, the effect size is on average 0.050 g (95% CI: 0.026 to 0.076) higher. Moreover, studies comparing foster care children with children in the majority population report on average 0.493 g (95% CI: 0.293 to 0.724) higher effect than longitudinal studies without comparison groups. Studies using other measurement instruments (see table 2) report on average 0.308 g (95% CI: -0.517 to -0.099) lower effect compared to studies using CBCL. These effect moderators explain about 23 percent of the heterogeneity between the included studies ($I^2_{\text{adjusted}} = 23.25\%$).

In family service-oriented regimes, the results show that studies reporting on cognitive functioning has on average 0.530 g (95% CI: 0.874 to 0.186) lower effect than studies reporting on adaptive functioning. The publication year of the study is also significantly associated with the effect of care. For every increase in publication year, the effect size is on average 0.028 g (95% CI: 0.014 to 0.043) higher. These effect moderators explained close to 24 percent of the heterogeneity between the included studies ($I^2_{\text{adjusted}} = 23.94\%$).
Table 7
The association between study-related covariates and effect estimates, results from random-effects meta-regression analysis

<table>
<thead>
<tr>
<th>Variable</th>
<th>Child protection-oriented welfare regimes&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Family service-oriented welfare regimes&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crude (95% CI)</td>
<td>Adjusted (95% CI)</td>
</tr>
<tr>
<td><strong>Study characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attrition</td>
<td>-0.004 (-0.004 to 0.003)</td>
<td>-</td>
</tr>
<tr>
<td>Length of follow-up in months</td>
<td>-0.001 (-0.006 to 0.004)</td>
<td>-</td>
</tr>
<tr>
<td>Mean age&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.032 (0.005 to 0.057)*</td>
<td>0.050 (0.026 to 0.076)*</td>
</tr>
<tr>
<td>Percent females&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.001 (-0.004 to 0.007)</td>
<td>-</td>
</tr>
<tr>
<td>Sample size</td>
<td>-0.000 (-0.001 to 0.000)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Study methodology and design</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No comparison group</td>
<td>-0.150 (-0.317 to 0.018)</td>
<td>(reference)</td>
</tr>
<tr>
<td>Children staying at home</td>
<td>-0.103 (-0.312 to 0.107)</td>
<td>0.232 (-0.028 to 0.491)</td>
</tr>
<tr>
<td>Children in majority population</td>
<td>0.294 (0.102 to 0.485)*</td>
<td>0.493 (0.263 to 0.724)*</td>
</tr>
<tr>
<td><strong>Outcome type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptive functioning</td>
<td>-0.104 (-0.298 to 0.090)</td>
<td>(reference)</td>
</tr>
<tr>
<td>Cognitive functioning</td>
<td>-0.077 (-0.312 to 0.157)</td>
<td>-</td>
</tr>
<tr>
<td>Internalizing behavior</td>
<td>-0.380 (-0.762 to 0.002)</td>
<td>-</td>
</tr>
<tr>
<td>Externalizing behavior</td>
<td>0.134 (-0.060 to 0.329)</td>
<td>-</td>
</tr>
<tr>
<td>Total behavior problems</td>
<td>0.106 (-0.076 to 0.287)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Measurement instrument</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBCL&lt;sup&gt;e&lt;/sup&gt;</td>
<td>0.106 (-0.076 to 0.287)</td>
<td>(reference)</td>
</tr>
<tr>
<td>CBCL + Other</td>
<td>0.044 (-0.155 to 0.244)</td>
<td>-0.141 (-0.390 to 0.108)</td>
</tr>
<tr>
<td>TRF&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.566 (0.267 to 0.865)*</td>
<td>0.061 (-0.304 to 0.427)</td>
</tr>
<tr>
<td>Other&lt;sup&gt;g&lt;/sup&gt;</td>
<td>-0.284 (-0.462 to -0.107)*</td>
<td>-0.308 (-0.517 to -0.099)*</td>
</tr>
</tbody>
</table>
### External factors

<table>
<thead>
<tr>
<th></th>
<th>Peer-reviewed (ref = non-peer-reviewed)</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.163 (-0.359 to 0.033)</td>
<td>0.002 (-0.012 to 0.007)</td>
</tr>
<tr>
<td>Percent heterogeneity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>explained</td>
<td>tau² = 0.198, η² = 99.90%,</td>
<td>tau² = 0.1132, η² = 98.48%</td>
</tr>
<tr>
<td></td>
<td>Fadjusted = 23.25%</td>
<td>Fadjusted = 23.94%</td>
</tr>
</tbody>
</table>

- kcrude = 152; kadjusted = 124; b kcrude = 87; kadjusted = 87; c Child protection oriented: kcrude = 124; kadjusted = 124; d kcrude = 147; kcrude = 85; e CBCL = Child Behavior Check List; f TRF = Teacher’s Report Form; g See Table 2; * p<.05
Discussion

Taking on a meta-analytical approach, this thesis aimed to further our understanding regarding potential sources of heterogeneity between studies of mental health outcomes in children experiencing foster care. This was achieved by examining whether and to what extent various hypothesized sources of heterogeneity were linked to variation in effect sizes from 240 studies that have addressed whether foster care improves mental health-related outcomes.

The meta-analyses showed no significant and clinically relevant effect from foster care on adaptive functioning, cognitive functioning, externalizing behaviour, internalizing behaviour, or total problem behaviour. But the analysis also showed that there was considerable heterogeneity between the included studies, which was explored by means of meta-regression.

After relevant adjustments, it was shown that studies that had used the majority population as a comparison group had reported on average larger effect sizes compared to studies that had used similar children staying at home as a comparison group, and longitudinal studies without a comparison group. This finding was not expected, but the result might reflect a potential collinearity with an unknown covariate, which is a suitable topic for further research. Given the wide range of measurement instruments used by the studies, the significant unadjusted estimates for certain instruments motivate further research into this potential source of heterogeneity even though the adjusted estimate was insignificant. Looking at the evidence so far, it cannot be ruled out that different instruments might yield systematically different results. Both these results clearly bring attention to the need of standardizing the design and procedures of studies evaluating the effect of foster care on mental health outcomes.

As mentioned earlier, children placed in their teens tend to have poorer outcomes than younger children, but the mean age of study participants was not significantly associated with the effect of foster care in this study, despite that the covariate ranged from infants to adolescents. This result did not confirm indications from earlier studies. The percent of females in the study sample was not associated with the effect size either, even though some samples had no female study participants at all. The loss of study participants to follow-up did not seem to matter for the effect estimate. Effect estimates published in peer-reviewed journals were not significantly larger compared to those abstracted from non-peer reviewed publications.

This thesis also confirmed earlier results reported by Gypen and colleagues (2017), indicating no differences in mental-health outcomes between former foster care children in child protection-oriented and family service-oriented child welfare regimes. But the analysis was
also stratified according to child welfare regime in order to examine if the strength and pattern of effect moderators were the same. In child protection-oriented regimes, the mean age of the foster care child was significantly associated with larger effect sizes, and so were studies using children in the majority population as a comparison group. But studies that had used other measurement instruments than CBCL or TRF reported on average smaller effect sizes. This pattern of effect moderators was not found in family service-oriented child welfare regimes, instead, the type of mental-health outcome and publication year were significantly associated with larger effect sizes. This might indicate that the foster care practice of family service-oriented regimes has been improved over time, in contrast to the practice of child protection-oriented regimes. The amount of heterogeneity explained was more than double in these two meta-regression models compared to the non-stratified analysis. This indicates that it is meaningful to continue to search for potential sources of heterogeneity in different child welfare regimes.

An explanation to the different patterns of effect moderators might be sought in the different cultural, political and social contexts surrounding the foster care practice in the respective regimes. For example, children from minorities are overrepresented in the foster care systems of both Nordic (e.g. Franzén et al., 2008), and Anglo-Saxon countries (e.g. Chabot et al. 2013; Fallon et al. 2013; Fluke et al. 2010), but the processes of discrimination might affect different minorities in different ways as it depends on context specific traditions, policies, ideas, and practices. The political underpinnings of foster care are also changing over time, resulting in different ideas on which children to remove from home, and what to offer them while in care (Lundström, 1993; Parton, 2014). Thus, foster care children face different attitudes, assumptions, ideas, and practices while in care, which in turn may affect their mental development as they transform into adults. But within the scope of this thesis, it was not possible to investigate this further as most Nordic and Central European studies do not report on racial disparities within the sample as often as Anglo-Saxon studies. So, how various contexts might affect the outcome of foster care is also suggested as a suitable topic for further research.

The child protection practice has been strongly criticized all over the world, however, policymakers have recently made several attempts to improve the practice, so it is not far-fetched to assume that it has been substantially improved during the last decade. For this reason, it would have been interesting from a theoretical point of view to empirically test whether studies conducted before the 2000s have reported on average poorer outcomes than
studies conducted since the year 2000. In addition, it would have been interesting to compare older studies conducted in child protection-oriented child welfare regimes with older studies conducted in family service-oriented regimes. Unfortunately, this thesis has no access to data on when a study was conducted, but if such data were to be collected this topic could be examined by adding an interaction term between publication date and child welfare regime. To be more specific, further research could test if there is any additional negative effect from being placed during the 90s in a child protection-oriented regime, a topic of increasing relevance as policymakers are taking steps toward striking a new balance between child protection and family service (Gilbert et al., 2011; Östberg, 2012). Thus, they would be greatly helped if it can be shown that any of the child welfare regimes has been more successful in promoting the mental-health of foster care children.

**Strengths and weaknesses**

The foremost strength of this thesis was that enough effect sizes were collected to yield sufficiently precise estimates in the meta-regression analysis. Needless to say, the searching of databases, assessing the eligibility of the studies found, and abstracting necessary information from hundreds of studies is indeed a time-consuming endeavour. Another strength was the narrow inclusion criteria applied, which ensured a quite easy and straightforward interpretation of the summary effects.

Bergmark (2006) has summarized three major points of criticism concerning meta-analysis of relevance for this thesis. First, the ‘garbage in, garbage out’ problem refers to summary effect sizes being compromised by the inclusion of too many low-quality studies. Second, the ‘file drawer’ problem is also raised, i.e. that scientific journals prefer to publish statistically significant results, which may skew the summary effect estimate. Third, the ‘apples and oranges’ problem refers to situations in which interventions are too different for their respective effects to be reliably synthesized, a typical problem in both public health and social work as interventions are implemented differently due to the varying contexts of countries, regions and local areas. This thesis can be said to suffer from the ‘apples and oranges’ problems as it has synthesized effect sizes originating from four decades of research from varying cultural, social, and political contexts including Scandinavia, the Middle East, and various parts of the Anglo-Saxon world. The question is whether results from Iraqi Kurdistan can be straightforwardly synthesized with results from Scotland and Norway? Some researchers have also pointed out that (1) foster care is not as specific as many medical treatments, and that (2) the
diversity of the global population of foster care children is problematic because a certain degree of homogeneity in the studied population is assumed. These problems have caused some researchers to even question the ability of randomized control trials to isolate the intervention effect from other factors that affect study participants (Bergmark & Lundström, 2007; Bergmark et al., 2011).

The meta-regression analysis of this thesis relied solely on study-related covariates. It would have been more suitable to use covariates related to the individual foster care children. This typically leads to problems with internal validity. As an example, the covariate for age does not measure the actual age of the individual foster care children, but rather a mean value of the age of the foster care children in the sample. Such a value does not account for potential variation within study samples, hence, two study samples with the same mean age could indeed be very different. Thus, even though the estimate for mean age was insignificant in this thesis, it cannot be ruled that foster care children’s age actually is an important effect moderator. Previously, it was also mentioned that publication year cannot be used to reliably capture varying cultural, political, and social contexts in different times as it may differ too much from when the studies actually had been conducted. Moreover, this thesis has synthesized results from a wide range of measurement instruments without being able to ensure that these instruments actually measured the same thing, and the number of included studies was too small for allowing a more in-depth analysis of the variance in effect sizes attributable to the choice of instrument.

When working with quantitative data concerning human beings, important qualitative perspectives are typically lost, such as the foster care children’s own viewpoints on their mental-health development (cf. Silverman, 2013).

Conclusions

The results imply the need to standardize evaluations of the effect of foster care and the measurement of various mental-health outcomes. It is not enough to conduct randomized controlled trials, it is also important to conduct them in the same way to ensure reliable and synthesizable effect estimates. The need to standardize evaluations is made even more evident given the different patterns of heterogeneity between studies conducted in child protection-oriented and family service-oriented child welfare regimes. The foster care practice seems to have been improved over time in the family service-oriented countries, but before any clear implications for policymakers in child protection-oriented regimes can be drawn,
research need uncover whether the results reflect real differences in the effect of care, or any underlying factor unaccounted for in this analysis.
References


StataCorp (2017). *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.


