INTERNATIONAL JOURNAL OF RESEARCH & METHOD IN EDUCATION
https://doi.org/10.1080/1743727X.2023.2215699

Routledge
Taylor & Francis Group

Examining heterogeneity of education intervention effects using quantile mixed models: a re-analysis of a cluster-randomized controlled trial of a fluency-based mathematics intervention

Paul Thompson [a], Kaydee Owen[b] and Richard P. Hastings[a,c]

[a]Centre for Educational Development, Appraisal and Research (CEDAR), University of Warwick, Coventry, UK; [b]School of Educational Sciences, Bangor University, Bangor, UK; [c]Department of Psychiatry, Monash University, Clayton, Australia

ABSTRACT

Traditionally, cluster randomized controlled trials are analyzed with the average intervention effect of interest. However, in populations that contain higher degrees of heterogeneity or variation may differ across different values of a covariate, which may not be optimal. Within education and social science contexts, exploring the variation in magnitude of treatment effect at different points in the population can indicate where the intervention is most effective rather than assuming an average effect.

Data from [Owen, K.L., et al., 2021. Implementation support improves outcomes of a fluency-based mathematics strategy: A cluster-randomized controlled trial. *Journal of research on educational effectiveness*, 14 (3), 523–542.] were reanalyzed using three modeling approaches: conditional mean-modeling reporting the average treatment effect using linear mixed models, and two quantile regression-based methods. Quantile regressions report the quantile treatment effects at different percentiles: 10th, 25th, 50th, 75th and 90th. For the Quantile approaches, a significant intervention effect in the median to upper quantiles was found and linear quantile mixed model showed improved fit over the other approaches.

An improved picture of intervention effects may be apparent using quantile regression methods when analyzing cluster randomized trials that have heterogeneous error variance. In particular, the linear quantile mixed model shows improved model fit allowing a multilevel framework to include random effects. There is considerable scope to extend this framework to incorporate more complex RCT designs.

ARTICLE HISTORY
Received 25 July 2022
Accepted 1 April 2023

KEYWORDS
RCT; quantile regression; linear quantile mixed models; general linear mixed models; fluency-based mathematics

Introduction

In most randomized controlled trials (RCTs), the primary driver of the study is to determine whether the average treatment effect is different from zero or an improvement on an existing intervention or standard practice. The average treatment effect is clearly an appealing summary to show the comparative effectiveness of an intervention, but the presence of heterogeneity of the treatment effect remains under-investigated when only considering the average (Angus & Chang, 2021). When working with populations with a high degree of variability in their characteristics, the average

CONTACT Paul Thompson paul.thompson.2@warwick.ac.uk Centre for Educational Development, Appraisal and Research (CEDAR), University of Warwick, Coventry, CV4 7AL, UK

may only provide a limited glimpse of the effectiveness of an intervention (Koenker *et al*., 2017). One solution is to perform sub-group analyzes but, unless designed into the study, these are often underpowered and exploratory with higher potential for type II error (Burke *et al*., 2015). In addition, if the trial has been appropriately randomized, the heterogeneity of treatment effects will still be present as the control and intervention groups should follow similar empirical distributions (Gabler *et al*., 2009). More specifically, if factors included in the randomization do not account for the heterogeneity, then the heterogeneity issue persists in each arm of the trial. This is a particular issue in pragmatic trials (Giraudeau *et al*., 2022).

|Focusing on average effects of intervention at a policy level in education and social care may not be the optimal strategy (Huber & Wüthrich, 2019, Hohberg, Pütz, & Kneib, 2020). The average effect could be negligible, but there may be substantial benefits at the bottom of the distribution. In such a situation, the intervention may be especially useful for reducing inequalities. For example, suppose we have two programs focusing on numeracy skills and one shows improvements only in the higher performing children, whereas the second substantially benefits children with weaker numeracy skills. Policy makers may have a particular interest in the latter finding since it could offer an opportunity to reduce educational inequalities. However, such conclusions may be drawn typically via underpowered sub-group analysis (Dijkman, Kooistra, & Bhandari, 2009, Burke *et al*., 2015, Tang *et al*., 2021).

The average treatment effect calculation may underestimate an effect or even miss differential effects of sub-groups entirely (Angus & Chang, 2021). Traditional methods for estimating interventions are usually from medical trials where, arguably, the populations are more homogeneous by design – with strict inclusion and exclusion criteria to limit variability. It is crucial to understand the pragmatic relevance of heterogeneity when considering analysis methods for educational and social care trials, as certain interventions may be more beneficial to certain groups within the population (Hohberg *et al*., 2020). For example, when using cluster randomization for school-based interventions, the interest may not be in finding an average effect but understanding the distribution of treatment effect at all student ability levels, with specific interest in those who are under-performing.

An alternative approach to using average treatment effects is to report quantile treatment effects (Koenker *et al*., 2017). The quantile treatment effect (QTE) is the difference between particular quantiles of the outcome distribution in the intervention group and the outcome distribution in the control group (Callaway, 2019). This effectively provides an estimate of the treatment effect at different points in the outcome distribution. For example, we might find larger treatment effects at the extreme ends of the population, but more modest effects at the median population area. Several accessible introductions to methods of analysis to evaluate the average treatment effect are available in the RCT literature (see for example: Altman, 1991, Twisk *et al*., 2018). However, accessible guides are less frequently provided for quantile treatment effects.

Konstantopoulos *et al*. (2019) presented a general overview of quantile regression to estimate intervention effects, and some researchers reporting RCT results using QTEs illustrate the potential value of the approach. For example, Ohrnberger *et al*., (2020) discussed the effects of conditional cash transfer programmes on mental health, finding substantial variation in the magnitude of effect of the programme across the mental health distribution. Those with the worst mental health profile showed a positive intervention effect approximately four times the average effect. Chen *et al*. (2016) presented results from two trials of interventions in Alzheimer's disease showing that treatment effects appeared larger at the higher percentiles of progression but less at the lowest percentiles.

In data with no clustering, (i.e. dependency of observations, for example, individuals within the same school will potentially respond more similarly then their peers from another school), quantile treatment effects can be estimated using quantile regression. Standard linear regression models the average relationship between dependent and independent variables, whereas quantile regression models the relationship at various sample quantiles (Yu, Lu, & Stander, 2003). Quantile regression has some additional benefits over ordinary least squares regression. In particular, there are no

distributional assumptions for quantile regression, so the approach is more robust to outliers, skew, and multimodality (Koenker, 2005).

The quantile treatment effect in the presence of clustering with the data is less frequently discussed in the literature. Konstantopoulos *et al*. (2019) discussed the application of quantile regression more broadly in randomized controlled trials in education and presented a specific example of a trial with clustered data. Their approach to dealing with the clustered nature of the data was to use corrected standard errors rather than modeling the dependence directly (Parente & Santos-Silva, 2016). Hagemann (2017) provided an improved procedure of cluster corrected standard errors using a wild gradient bootstrap procedure when the number of clusters is small (a common issue in other correction procedures). The comparison of cluster corrected standard errors versus multilevel models is relatively under-explored. Cheah (2009) made direct comparison between models calculating the average treatment effect and, more informally, Gelman (2009) briefly discussed a preference for the multilevel approach in a blogpost. However, to our knowledge, no direct comparison of quantile regression models with clustered data using either cluster corrected standard errors or multilevel models, has been reported to date.

In this paper, we discuss and make direct comparison between two approaches to estimation of quantile treatment effects when the data are clustered and compare the results of the quantile treatment effects with original analyses reporting the average treatment effect in a cluster RCT in education (Owen *et al*., 2021).

Methods

We summarize the trial reported by Owen *et al*. (2021). More detailed descriptions are discussed by Owen *et al*. (2021), but we cover the essential elements required to understand the trial when considering the implications for comparison of statistical approaches.

Trial design and participants

The trial was delivered to schools across North Wales, and within six counties (Conwy, Denbighshire, Flintshire, Gwynedd, Anglesey, and Wrexham). Nominated teachers from each school attended a 3-hour training session for the mathematics intervention, Say-All-Fast-Minute-Every-Day-Shuffled (SAFMEDS). This training was delivered prior to randomization to trial arm (intervention plus support vs intervention with no support), so that bias was minimized, and trial differences were subject to chance. In addition, each school selected up to 10 children to participate in the trial.

Schools in the North Wales region responded to an open advertisement for this project. Upon expression of interest, the researchers asked them to identify children needing intervention support to master, and build fluency in, basic arithmetic skills. Due to the children being in different phases of their schooling, the researchers asked schools to identify children based on the following criteria:

- 6–7 years olds (in Year 2 classes) who were working below the expected standard for their age in mathematics and numeracy for their age required an intervention support. At this age, children in Wales are not formally assessed using standardized measures, so the researchers allowed teachers to make their own judgements based on experience and class test data.
- Children $\geq$7 years old who scored less than 100 standard points on the national numeracy procedural test undertaken at the end of the preceding academic year.

The majority children in the sample were aged 6–8 years. For supporting context, the two secondary schools who responded to the advertisement represent, (i) a school with Year 7 children (aged 11–12 years) who significantly underperformed on the procedural test, (ii) a special educational

needs school that supported children aged 11–17 years working below the age-expected level in numeracy/mathematics.

The mean age of the children attending schools in the no ongoing support trial arm was 7-years 3-months (*SD* = 14.34 months, range: 6-years 0-months to 9-years 2-months). The mean age across the ongoing support trial arm (intervention) was also 7-years 3-months (*SD* = 14.32 months; range: 6-years 0-months to 15-years 10-months). Consent was obtained for 575 children (N[Support] = 294, N [NoSupport] = 281), across 60 schools (N[Support] = 31, N[NoSupport] = 29). A full summary tables of baseline characteristics of schools and children can be found in Owen *et al*. (2021).

Three other measures were recorded for use as control variables in the analysis. These include:

Predominant home language – As this study is conducted in Wales, some children may predominantly speak Welsh as their home language with English as a second language, or vice versa.

Eligibility for free school meals (eFSM) – Free school meals are available in Wales for children from families typically with lower incomes or in receipt of social security benefits. We gathered data on whether each child was eligible to receive free school meals based on Welsh government criteria in place at the time of the study.

Gender – reported gender of the child (male/female).

Randomization

Randomization was conducted using minimization (Altman & Bland, 2005, Kahan & Morris, 2012) and schools were allocated to one of the two trial arms (intervention plus support vs intervention with no support). The allocation was stratified by county and the language used predominantly for teaching in the school (either English or Welsh). All teachers received training for the SAFMEDS intervention prior to the children completing the baseline assessments.

Intervention

All teachers were given training in the Say-All-Fast-Minute-Every-Day-Shuffled (SAFMEDS) strategy which aims to improve children's fluency of basic mathematics skills (Tyler *et al*., 2018). Owen *et al*. (2021) focused on randomizing schools to SAFMEDS with and without ongoing support to examine whether coaching support would improve children's numeracy outcomes (by improving fidelity of the teacher's implementation). The ongoing support arm of the trial received three in-situ support visits from an experienced researcher with several years of experience of using SAFMEDS strategies in schools. The no support arm received the same initial training in SAFMEDS strategy at baseline, but then received no further support from the experienced researcher, with the exception of technical issues or data input.

Analysis

The data were analysed using three different statistical methods for comparison to the original study: General linear mixed model, quantile regression using robust standard errors and linear quantile mixed models. The original trial (Owen *et al*., 2021) was analysed using a linear mixed effects model and found a significant interaction between time and intervention which was the effect of interest (three level model: Level 1 = time, level 2 = Children, level 3 = school). In this reanalysis, we adjust the model specification to include two hierarchical levels, Children were nested within schools (level 1 = children, level 2 = school). We opted to reframe the original model to allow direct comparison to the existing quantile regression approaches that permit two-level models. To the authors knowledge, an extension to a three-level quantile mixed models has not yet been developed but is mentioned later in our discussion as a potential future development.

Time is not incorporated into the models' setup. Instead, we choose to use baseline outcome as a covariate in the model and the dependent measure becomes the outcome at study endpoint follow-

up. Either method of analysis is permitted to assess whether the intervention shows an effect (O'Connell *et al*., 2017). The effect of interest in this design is the main effect of intervention.

All models adjust for the multilevel nature of the data using a linear mixed effects model structure or model with adjusted robust standard errors to account for clustering in the data. Similarly, all models adjust for the same covariates in the model and in this respect are directly comparable. Linear mixed-effect models partition the variance of the outcome variable into component levels of the hierarchy (Gelman & Hill, 2007, Galecki & Burzykowski, 2013), . To correspond with the original analysis from Owen *et al*. (2021), children's raw scores on the Mathematics Fluency and Calculation Tests (MFaCTs) measures were used as standardized scores were not available.

The linear mixed model (LMM) including the intervention indicator but without covariates can be written as follows,

$$\begin{aligned} level1{:}Y_{ij} &= \beta_{0j} + \beta_{1j} Baseline_{ij} + R_{ij} \\ level2{:}\beta_{0j} &= \gamma_{00} + \gamma_{01} Intervention_j + U_{0j} \\ \beta_{1j} &= \gamma_{10} \end{aligned}$$

where β are the parameter estimates, U_{0j} is the random intercept for schools, γ_{01} is the intervention main effect, and γ_{10} is the baseline outcome. Level 1 within the LMM models contained covariates associated with individual children: Baseline outcome, gender, predominant home language, eligibility for free school meals status (eFSM), and school year group. At level 2 (school), covariates included: school administrative county and trial arm.

The remaining two methods estimate the quantile treatment effects at the sample quantiles 10th, 25th, 50th, 75th and 90th. Essentially, the choice of quantile for the purposes of demonstration of the method is somewhat arbitrary. Any quantile across the distribution could have been specified but for brevity, we chose a limited selection spanning the range of the distribution. The first quantile-based approach used linear quantile mixed models proposed by Geraci and Bottai (2014) and incorporated a multilevel structure including random effects. A second quantile approach was included for comparison fitting the quantile regressions without random effects, but instead reported robust standard errors (Parente & Santos-Silva, 2016, Konstantopoulos *et al*., 2019), .

Standard ordinary least squares regression is estimated via minimization of the sum of squares with respect to the parameters. For example, the regression can be written as follows for an individual *i*,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where y is the outcome variable, x is the predictor variable, ϵ is the residual (error term) of y and β are the mean regression parameters. Typically, to estimate the parameter vector $\boldsymbol{\beta}$, via the quadratic loss function, $r(u) = u^2$, given a data set of observations $\{\boldsymbol{x_i}, y_i\}_{i=1}^n$ and involves minimization of the sum of squared residuals as follows (Yu, Lu, & Stander, 2003),

$$\min \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2,$$

In contrast, quantile regression is estimated by minimization of the weighted sum of the absolute values of the residuals for quantile q (Koenker, 2005, Koenker *et al*., 2017). For each sample quantile, q, we estimated a corresponding set of parameter estimates. The corresponding linear regression for each quantile can be written as follows,

$$y_i = \beta_0^q + \beta_1^q x_i + \epsilon_i$$

where q indicates the quantile of interest $[0 < q < 1]$. The corresponding loss function for quantile regression is the absolute value, $r(u) = |u|$ or more commonly written according to a specific quantile, τ, $\rho_\tau(u) = \tau|u|$ Furthermore, the parameter estimates from the quantile regression model are

estimated by minimizing the weighted sum of the absolute values of the residuals for quantile, τ (Yu *et al*., 2003, Konstantopoulos *et al*., 2019), ,

$$argmin\left[\sum_{i=1}^{N}\rho_{\tau}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\right],$$

where ρ_τ is the check function, $min\left[\tau\sum_{i=1}^{N}|y_i - \boldsymbol{x}_i^T\boldsymbol{\beta^{\tau}}| + (1-\tau)\sum_{i=1}^{N}|y_i - \boldsymbol{x}_i^T\boldsymbol{\beta^{\tau}}|\right]$, and τ is a quantile of interest. Moving from a simpler quantile regression model to a linear quantile mixed model which allows the inclusion of random effects was proposed by Geraci and Bottai (2014). The linear quantile mixed model follows the same general framework to a standard linear mixed model and all parameters are τ-dependent. The model with a single random effect can be written as follows,

$$y_{ij} = x_{ij}^T\theta_x^{(\tau)} + z_{ij}^T u_i + \boldsymbol{\epsilon}_{ij}^{(\tau)}$$

where $\theta_x^{(\tau)}$is a vector of unknown fixed effects, τ is the quantile of interest, $u_i = (u_{i1}, \ldots, u_{iq})$ are the zero-median vector of random effects which are independent of the model error term $\boldsymbol{\epsilon}_{ij}^{(\tau)}$, $i = 1, \ldots, M$ and conditional on the $\Psi^{(\tau)}$ covariance matrix (q x q). Hence, the random effects, u_i are also dependent on τ via $\Psi^{(\tau)}$. Geraci & Bottai (2014) define the *i*th contribution to the marginal likelihood by integrating out the random effects, $L_i(\theta_x, \sigma, \Psi|y_i) = \int_{R^q} p(y_i, u_i|\theta_x, \sigma, \Psi)du_i$,

where $R^q$is the q-dimensional Euclidean space.

All analyses use R version 4.0.3 (2020-10-10), and R packages: linear mixed models for average treatment effect estimation uses 'lme4'; quantile regression using robust standard errors, 'quantreg' (Koenker, 2021); and linear quantile mixed models, 'lqmm' (Geraci, 2014). The analysis R code is available at the Open Science Framework project page (https://osf.io/jqukn/).

Results

The data from Owen *et al*. (2021) were reanalyzed using a linear mixed model to measure the average treatment effect (ATE) and comparisons made with two alternative quantile regression analysis approaches that focused on quantile treatment effects. The primary outcome data were skewed and contained non-negative integers (see Figure 1 for both school Grades 1–2 and Grades 3–5 assessments), so model assumptions for the linear mixed model were checked and heterogeneity of variance was observed (see Appendix 1). Figure 1 highlights the skew in the data indicated by longer right tails in the distributions, seen in each panel when the shape of each distribution is not reflected around the central tendency (median, horizontal bar). When fitting the adjusted linear mixed model with baseline covariate adjustment, we found that the intervention was not statistically significant ($\beta = 2.81$, $p = .061$) which differed from the reported results from Owen *et al*. (2021). Important to note is that the original Owen *et al*. (2021) model and the baseline adjusted model are not directly comparable as the model structure is different, but the test of the intervention is similar in principle. Given the limitation of the current quantile models to only permit two-levels, the two level baseline model for the average treatment effect provides the most suitable comparison to the new quantile methods. Table 1 shows the full model output for the ATE under the linear mixed model with baseline adjustment.

MFaCTs: grades 1–2 assessment

Table 1 presents the comparison of model parameter estimates, for the linear mixed effect model (ATE) and both quantile-based models reporting the QTEs for grades 1–2 assessment data. Despite the non-significance of the ATE, it appears from both quantile models that at some points in the distribution that a statistically significant effect of treatment effect is observed. In particular, and consistently with both quantile models, the 25th, 50th, 75th and 90th quantiles show varying levels of statistical significance. The largest observed effects are found using the linear

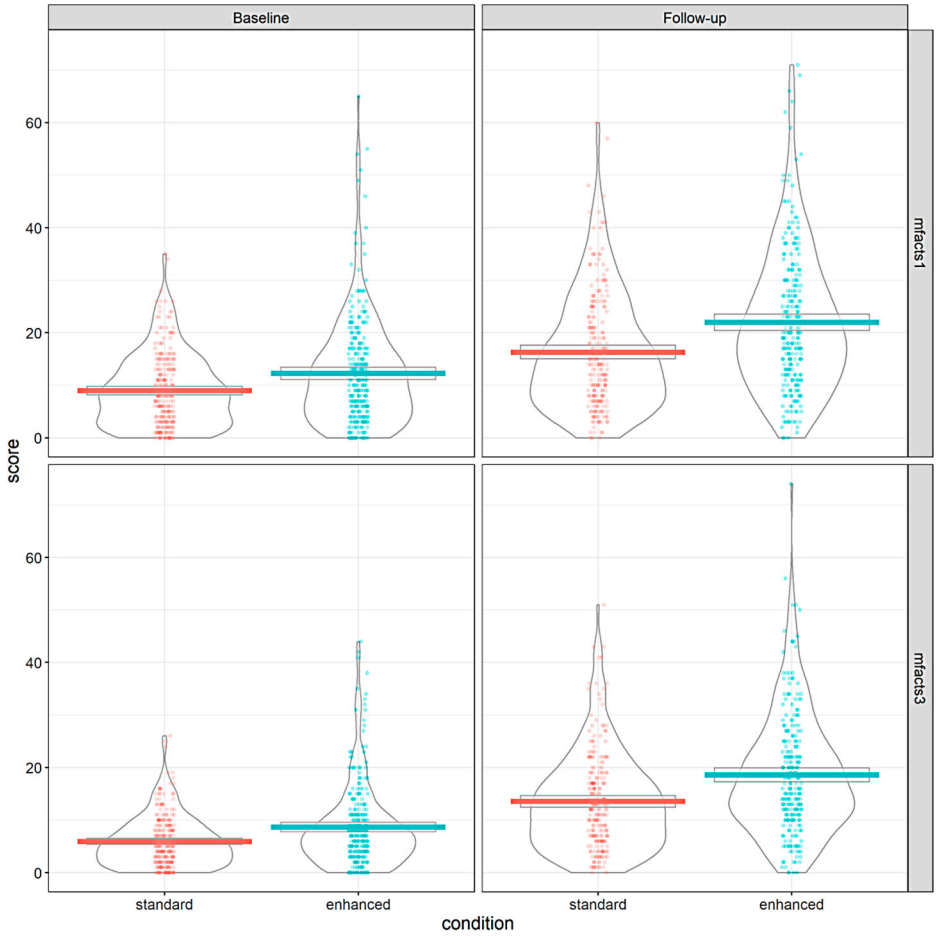

Figure 1. The pirate plots of the (Owen *et al*., 2021) data for Grades 1–2 and 3-5. The data are shown split by time point and trial arm.

quantile regression with robust standard errors ($p < 0.001$ at 25th and 50th quantiles; and $p < 0.01$ at 75th and 90th) compared to the effects estimated using the linear quantile mixed model ($p < 0.05$ at 25th, 50th, 75th, and 90th quantiles) which were more modest. Although control variables were not of primary interest – and should not be interpreted with substantial meaning – we did find that across the quantiles, certain control variables did show statistical significance, compared to the linear mixed model (ATE). For example, eligibility for free school meals status (eFSM) was found to show a consistent (both quantile models) and statistically significant effect in the upper quantiles (75th and 90th), but not at the lower quantiles.

Figure 2 shows the estimated coefficients at different percentiles from the quantile models and, for reference, the average treatment effect in the corresponding linear mixed model for MFaCTs: Grades 1–2. The red lines are the average treatment effect; the green lines are the linear quantile mixed model; and the blue lines are the linear quantile regression with robust SEs. In general, both quantile models show that the QTEs vary across the percentiles and are often relatively different magnitudes depending on their location in the outcome distribution. In these data, it is clear that the ATE does not adequately capture the full picture of effects in either intervention or control covariates as neither quantile coefficients are parallel to their ATE equivalents across the percentiles.

Table 1. The model outputs for all models across all fitted quantiles for MFacTs: Grades 1-2.

			10th				25th				50th				75th				90th			
	ATE: Owen *et al.* (2021)		Quantile Reg (robust SE)		Linear Quantile Mixed Model		Quantile Reg (robust SE)		Linear Quantile Mixed Model		Quantile Reg (robust SE)		Linear Quantile Mixed Model		Quantile Reg (robust SE)		Linear Quantile Mixed Model		Quantile Reg (robust SE)		Linear Quantile Mixed Model	
Coefficient	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E
Intercept	1.89	3.48	**4.41** ***	1.11	4.16	4.37	3.05	2.59	4.41	5.68	**6.04** *	3.02	4.95	5.02	**7.44** *	3.42	5.30	4.97	−1.67	5.36	5.47	5.10
Baseline	**0.97** ***	0.06	**0.56** ***	0.06	**0.63** ***	0.07	**0.81** ***	0.10	**0.81** ***	0.12	**0.99** ***	0.09	**0.98** ***	0.07	**1.25** ***	0.09	**1.12** ***	0.10	**1.25** ***	0.08	**1.25** ***	0.17
Gender	0.60	0.83	−0.17	0.59	−0.05	0.79	−0.12	0.81	0.16	0.87	1.36	1.00	0.84	0.82	0.86	1.08	1.05	0.76	2.00	1.75	0.95	0.92
Predominant Home Language	1.68	1.56	**−1.20** *	0.49	−1.02	1.31	0.23	0.83	−0.31	1.62	−0.36	1.25	0.53	1.46	0.31	1.03	0.99	1.60	0.67	1.46	1.50	2.04
eFSM	−1.74	0.98	−0.66	0.55	−1.47	0.94	−1.81	0.93	−2.09	1.07	−1.54	1.05	**−2.04** *	0.89	**−3.52** ***	0.94	**−2.64** **	0.92	**−4.50** *	2.29	−2.52 *	1.15
Year	1.03	0.62	0.49	0.26	0.06	1.05	−0.23	0.95	−0.11	1.40	0.32	0.74	0.42	1.41	0.53	1.27	0.94	1.61	**6.00** **	1.92	2.60	1.46
County	0.18	0.41	**−0.48** ***	0.14	−0.28	0.35	0.08	0.25	0.15	0.43	0.08	0.29	0.00	0.31	0.28	0.27	0.38	0.55	0.33	0.59	−0.04	0.72
Intervention	2.81	1.46	0.84	0.51	2.26	1.38	**3.85** ***	1.03	**3.10** *	1.33	**3.65** ***	1.01	**3.22** **	1.16	**3.36** **	1.04	3.35 *	1.44	**4.50** **	1.38	**3.79** *	1.88
Random Effects																						
σ^2	72.06																					
τ_{00}	19.37_{sch}				1.53				7.74				6.75				10.35				21.84	
ICC	0.21																					
N	56_{sch}				55_{sch}				55_{sch}				55_{sch}				55_{sch}				55_{sch}	
Observations	464		464		464		464		464		464		464		464		464		464		464	
Marginal R^2/ Conditional R^2	0.447/0.564		−0.438				0.135				0.434				0.270				−0.850			

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

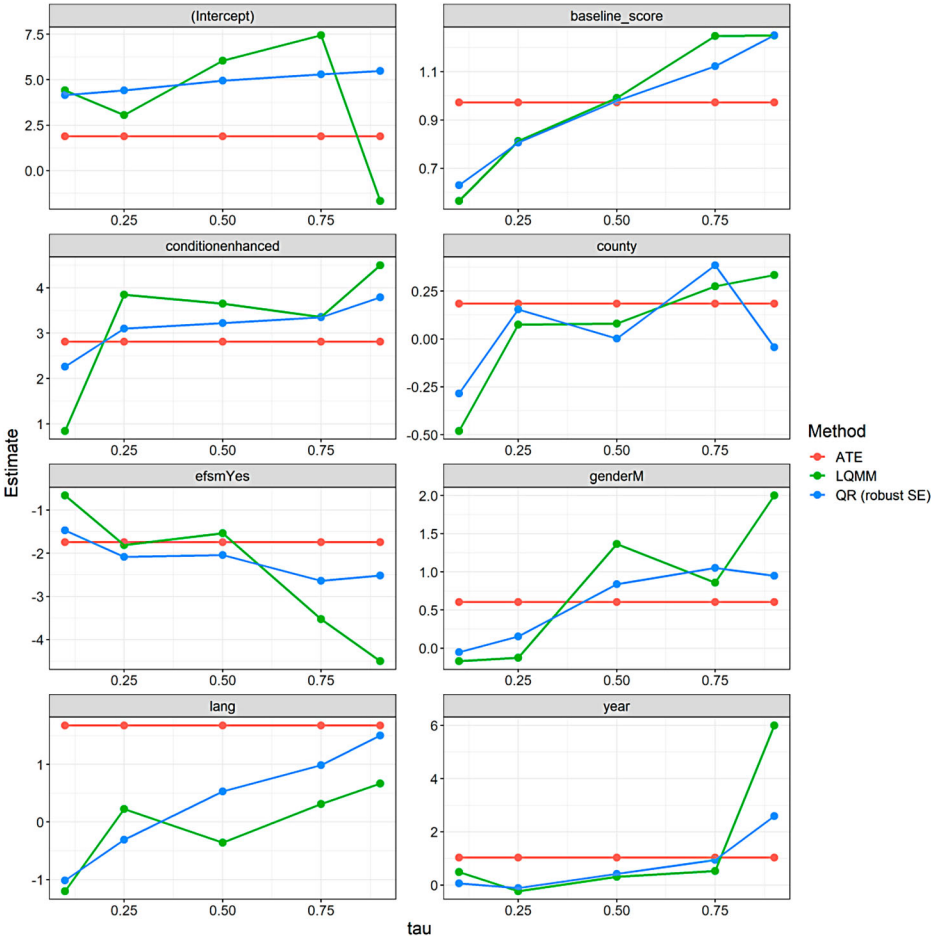

Figure 2. The estimated quantile regression coefficients at different percentiles and their corresponding ATEs from the linear mixed model for the MFaCTs: Grades 1–2 data.

Table 2 presents the comparison of model fit indices [AIC; Akaike (1973)] for both quantile models when fitted to the MFaCTs: Grades 1–2 data. The models are nested, so can be directly compared to each other using the fit indices. The linear mixed model fit indices are not included as they are not directly comparable to the quantile models given that they are fitted at different points in the distribution. The linear quantile mixed model shows a clear improvement in fit over the linear quantile regression with robust SEs as all AIC values are lower and at a magnitude that suggests that there is very strong evidence to favour the linear quantile mixed model.

Table 2. The AIC model fit indices for each quantile regression model at the 10th, 25th, 50th, 75th, and 90th quantiles for MFaCTs: Grades 1-2.

Model	10th	25th	50th	75th	90th
Linear Quantile Mixed Model	3,425.429	3,363.589	3,358.200	3,477.126	3,641.768
Quantile regression with robust SE	3,430.998	3,397.948	3,392.291	3,509.247	3,728.225

MFaCTs: grades 3–5 assessments

The data from the MFaCTs: Grades 3–5 assessment presents a slightly different set of results to the previous Grades 1–2. Table 3 presents the comparison of model parameter estimates, for the linear mixed effect model (ATE) and both quantile-based models reporting the QTEs for the Grades 3–5 assessments data. We find that the average treatment effect is again statistically non-significant compared with the results reported in Owen *et al*. (2021), but there is less consistency between the two quantile regression approaches. The linear quantile regression with robust SEs indicated statistically significant effects of intervention at the 25th, 50th and 75th quantiles ($\beta_{0.25} = 2.77$, $p = < .001$; $\beta_{0.5} = 2.46$, $p = .001$; and $\beta_{0.75} = 2.86$, $p = .008$ respectively). However, the linear quantile mixed model showed more modest effects for the intervention across the distribution but did additionally indicate that the 90th percentile showed differences ($\beta_{0.25} = 2.31$, $p = .064$; $\beta_{0.5} = 2.41$, $p = .033$; $\beta_{0.75} = 2.32$, $p = .012$; and $\beta_{0.90} = 2.87$, $p = .018$ respectively).

Figure 3 shows the estimated coefficients at different percentiles from the quantile models and, for reference, the average treatment effect in the corresponding linear mixed model for MFaCTs: Grades 3–5. The red lines are the average treatment effect; the green lines are the linear quantile mixed model; and the blue lines are the linear quantile regression with robust SEs. Again, both quantile models show that the QTEs vary across the percentiles and are often relatively different

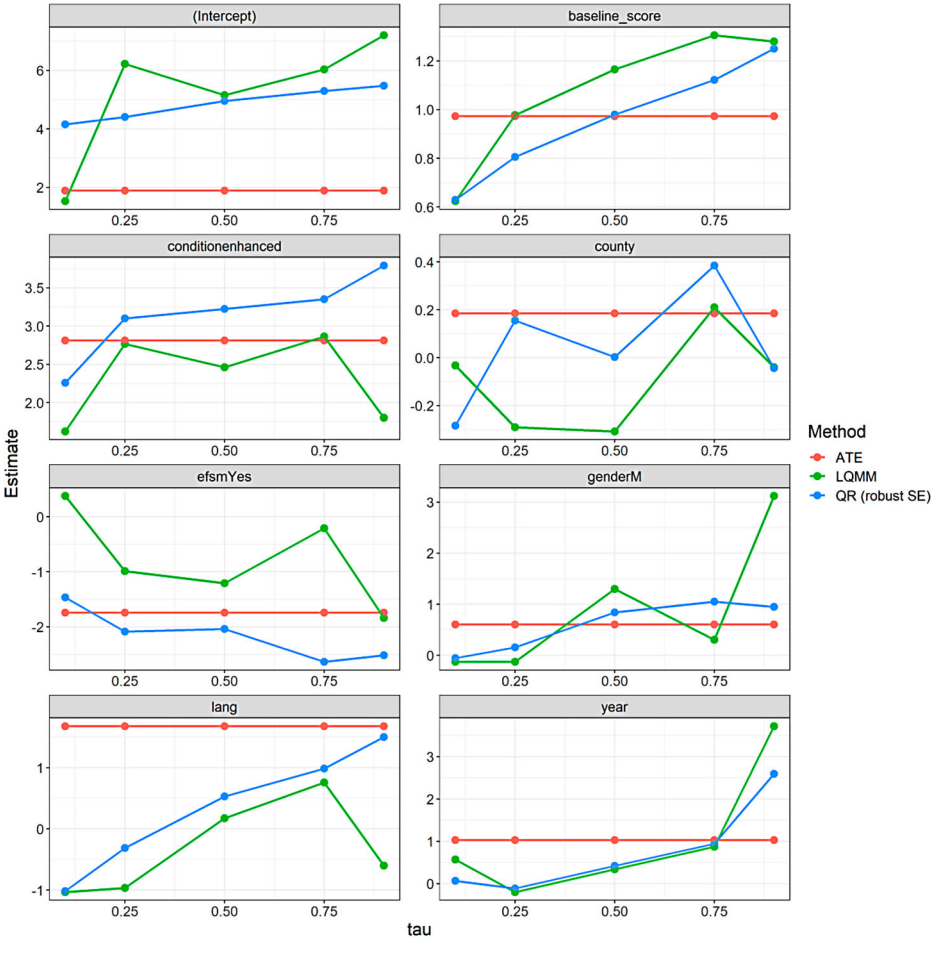

Figure 3. The estimated quantile regression coefficients at different percentiles and their corresponding ATEs from the linear mixed model for the Grade 3–5 data.

Table 3. The model outputs for all models across all fitted quantiles for MFaCTs: Grades 3-5.

			10th				25th				50th				75th				90th			
	ATE: Owen *et al.* (2021)		Quantile Reg (Robust SE)		Linear Quantile Mixed Model		Quantile Reg (Robust SE)		Linear Quantile Mixed Model		Quantile Reg (Robust SE)		Linear Quantile Mixed Model		Quantile Reg (Robust SE)		Linear Quantile Mixed Model		Quantile Reg (Robust SE)		Linear Quantile Mixed Model	
Coefficient	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E	Estimates	S.E
Intercept	5.59	3.06	1.53	2.66	6.57	5.19	**6.22** ***	1.81	7.09	5.06	**5.15** *	2.19	7.44	5.31	6.03	3.74	7.55	5.12	7.20	3.73	8.08	4.26
Baseline	**1.03** ***	0.07	**0.62** ***	0.16	**0.83** ***	0.13	**0.98** ***	0.08	**0.95** ***	0.08	**1.16** ***	0.10	**1.13** ***	0.11	**1.31** ***	0.13	**1.28** ***	0.14	**1.28** ***	0.15	**1.32** ***	0.16
Gender	1.18	0.73	−0.13	0.81	−0.15	0.76	−0.13	0.66	0.56	0.64	1.30	0.69	1.13	0.81	0.31	1.06	1.16	1.09	3.12	1.67	1.31	1.01
Predominant Home Language	−0.05	1.37	−1.03	0.73	−1.64	1.55	−0.97	0.79	−0.85	1.26	0.18	0.70	−0.31	1.38	0.76	1.39	0.14	1.38	−0.60	1.24	0.72	1.36
eFSM	−1.39	0.86	0.38	0.80	−0.85	0.91	−0.99	0.66	−1.23	0.92	−1.21	0.92	−1.35	1.20	−0.21	1.27	−1.56	1.02	−1.84	1.49	−1.64	1.07
Year	0.93	0.54	0.57	0.97	−0.39	1.42	−0.20	0.40	−0.28	1.41	0.34	0.78	0.03	1.54	0.87	1.04	0.66	1.93	**3.72** *	1.53	1.90	1.89
County	−0.11	0.36	−0.03	0.19	−0.20	0.29	−0.29	0.16	−0.40	0.30	−0.31	0.21	−0.27	0.31	0.21	0.37	0.27	0.45	−0.04	0.32	0.04	0.55
Intervention	2.31	1.29	1.62	1.13	1.84	1.11	**2.77** ***	0.65	2.31	1.30	**2.46** ***	0.74	**2.41** *	0.99	**2.86** **	1.08	**2.32** *	1.13	1.80	1.51	**2.87** **	1.04
Random Effects																						
σ^2	55.09																					
τ_{00}	15.15 $_{sch}$				7.15				4.44				4.53				15.28				17.34	
ICC	0.22																					
N	56 $_{sch}$				55 $_{sch}$				55 $_{sch}$				55 $_{sch}$				55 $_{sch}$				55 $_{sch}$	
Observations	460		460		460		460		460		460		460		460		460		460		460	
Marginal R^2/ Conditional R^2	0.411/0.538		−0.516				0.168				0.398				0.193				−0.777			

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table 4. The AIC model fit indices for each quantile regression model at the 10th, 25th, 50th, 75th, and 90th quantiles for grades 3-5.

Model	10th	25th	50th	75th	90th
Linear Quantile Mixed Model	3,245.461	3,194.719	3,207.521	3,310.57	3,457.473
Quantile Regression with robust SE	3,317.495	3,216.214	3,229.804	3,370.43	3,556.412

magnitudes depending on their location in the outcome distribution. In these data, it is clear that the ATE does not adequately capture the full picture of effects in either intervention or control covariates as neither quantile coefficients are parallel to their ATE equivalents across the percentiles.

Table 4 presents the comparison of model fit indices [AIC; Akaike (1973)] for both quantile models for MFaCTs: Grades 3-5. Similar to the Grades 1–2 results, the linear quantile mixed model shows a clear improvement in fit over the linear quantile regression with robust SEs as all AIC values are lower and at a magnitude that suggests that there is very strong evidence to favour the linear quantile mixed model.

Discussion

In this paper, we demonstrate through reanalysis of educational trial data the potential benefits of reporting quantile treatment effects rather than focusing on ATE. We have presented two different quantile regression methods that permit estimation of QTEs and showed the comparison to a typically reported ATE analysis using linear mixed models. Both linear quantile models' approaches do not have the same distributional assumptions as the linear mixed models, so have more flexibility to model outcomes with different distributional forms. Specifically, the standard errors when heterogeneity in error variance is present may be poorly estimated. In addition, the effects of different levels of covariates can be seen on the outcome at the different percentiles providing a more complete picture of the relationships among outcome and predictors (Koenker, 2005). This may show that certain covariates have different effects at different levels in the outcome distribution which would not be apparent through conventional average treatment effect approaches.

A further benefit is that a conventional sub-group analysis would require using a portion of the full dataset and reduction of the sample size, which may increase the chance of type I and II errors and a reduced ability to include all covariates. A quantile approach does not require reduction in the data as it fits the model at specified quantiles using the full dataset, so statistical power is not compromised. In addition, a quantile approach has reduced statistical assumptions so may in fact have improved statistical power (Tarr, 2012, Petscher & Logan, 2014, Howard, 2018). Formal sample size calculations could also be conducted via Monte Carlo simulation as the gold standard approach in more complex models (Kumle *et al*., 2021). Given that this was existing data, we omitted a formal calculation to avoid criticism in the same vein as Hoenig and Heisey (2001). Future work using simulations would be necessary to fully understand and incorporate the multilevel nature of these quantile models in a sample size calculation.

When re-analysing these data potential improvements to the analysis given the features of the primary outcome measures, MFaCTs at grades 1–2 and 3-5, were potentially possible. The primary outcome was bounded at zero, contained positive integer responses and was skewed, so applying a model technique that does not require certain distributional assumptions, and is less affected by outliers and extreme data, improves the estimates standard errors. It should be noted that if all model assumptions are met in linear mixed effects models, then quantile regressions may be less efficient (i.e. when the error distribution follows a normal distribution and without heteroscedasticity; Koenker & Bassett-Jr, 1978).

The reanalysis of the trial data from Owen *et al*. (2021) highlighted several potential improvements to standard practice in analysing cluster randomized trial data when heteroscedasticity of intervention effect is suspected. This is particularly apparent in populations that are largely

heterogeneous such as in educational contexts or in individuals with developmental disorders that may manifest highly variable range of abilities, needs, or comorbidity of conditions. The quantile regression approach permits a more flexible framework with reduced assumptions and removing focus from the average to explore factors that may be important determinants in distinguishing interventions that can be targeted at different sub-groups (Lê Cook & Manning, 2013).

In the Owen *et al*. (2021) report, the conclusion based on the data analysis was that adding in a coaching element to a numeracy intervention overall improved child numeracy outcomes. A key question is whether the current analyses make a practical difference; do they lead to different conclusions? The answer is clearly yes. Data at both grades 1–2 and 3–5 in the current analysis shows that intervention effect differs according to the point in the distribution at which it is assessed. Across both assessments, those children in the higher percentiles are benefiting from the intervention more than those at the lower percentiles. The current version of the intervention may not then help to reduce mathematical education inequalities (if the findings were replicated). Similarly, we also found that some control covariates differed in magnitude of effect across the different percentiles of the outcome distribution, in particular the status of eligibility for free school meals was highlighted as showing a statistically significant reduction in the outcome in the higher percentiles but not at the lower percentiles. This could potentially be informative for future intervention design as the intervention could be adapted to permit an adjusted version for different sub-groups of individuals (for example, varying amounts of contact time in the intervention).

When implementing quantile regression methods in cluster randomized trials consideration must be given to the dependence within clusters and how this can be incorporated into the analysis procedure. We present two methods to permit analysis of quantile treatment effects in cluster trials, with a subtle but key difference in how the variance at level 2 is incorporated into the model. The first method follows work by Konstantopoulos *et al*. (2019) who did not directly model the level 2 variance, but adjusted the model estimated standard errors to be robust. We have presented an alternative approach that directly models the level 2 variance and permits improved fit to the data. The model framework can also be expanded to include further levels, for example permitting nesting of children within classroom, and classroom within schools. This will be considered in future work as currently the statistical basis for more than two-level quantile mixed models does not exist despite being theoretically possible. Hence, we believe this provides further flexibility to researchers when analysing trials with more complex structures and with an interest in heterogeneous effects of intervention. In any study conducting confirmatory research, researchers should be encouraged to prespecify transparent analysis plans. Our recommendation for adopting the linear quantile mixed model method in RCT is conditional on the specific study design and population of interest. If researchers suspect a heterogeneous sample, then prespecification of the quantile method as a secondary analysis would advisable and retain the average treatment effect method as a primary analysis to permit consistency in standard RCT reporting and meta analyses across studies.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Paul Thompson http://orcid.org/0000-0001-9940-6913

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. *In:* E. Parzen, K. Tanabe, and G. Kitagawa, eds. *Selected papers of hirotugu akaike*. New York, NY: Springer New York, 199–213.

Altman, D.G., 1991. *Practical statistics for medical research*. London: Chapman; Hall.

Altman, D.G., and Bland, J.M., 2005. Treatment allocation by minimisation. *BMJ*, 330 (7495), 843. doi:10.1136/bmj.330.7495.843.

Angus, D.C., and Chang, C.H., 2021. Heterogeneity of treatment effect: Estimating how the effects of interventions vary across individuals. *JAMA*, 326 (22), 2312–2313. doi:10.1001/jama.2021.20552.

Burke, J.F., et al., 2015. Three simple rules to ensure reasonably credible subgroup analyses. *BMJ*, 351, doi:10.1136/bmj.h5651.

Callaway, B. 2019. *Qte: Quantile treatment effects*. https://CRAN.R-project.org/package = qte.

Cheah, B.C. 2009. Clustering standard errors or modeling multilevel data?

Chen, Y.F., et al., 2016. Quantile regression to characterize solanezumab effects in Alzheimer's disease trials. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 2, 192–198. doi:10.1016/j.trci.2016.07.005.

Dijkman, B., Kooistra, B., and Bhandari, M., 2009. How to work with a subgroup analysis. *Canadian Journal of Surgery*, 52, 515–522.

Gabler, N.B., et al., 2009. Dealing with heterogeneity of treatment effects: Is the literature up to the challenge? *Trials*, 10, doi:10.1186/1745-6215-10-43m.

Galecki, A., and Burzykowski, T., 2013. *Linear mixed-effects models using r: A step-by-step approach*. New York: Springer Publishing Company, Incorporated.

Gelman, A. 2009. MS Windows NT kernel description. https://statmodeling.stat.columbia.edu/2009/08/21/clustered_stand_1/.

Gelman, A., and Hill, J., 2007. *Data analysis using regression and multilevel/hierarchical models (Vol. Analytical methods for social research)*. New York: Cambridge University Press.

Geraci, M., 2014. Linear quantile mixed models: The lqmm package for laplace quantile regression. *Journal of Statistical Software*, 57 (13), 1–29. http://www.jstatsoft.org/v57/i13/.

Geraci, M., and Bottai, M., 2014. Linear quantile mixed models. *Statistics and Computing*, 24 (3), 461–479. doi:10.1007/s11222-013-9381-9.

Giraudeau, B., Caille, A., Eldridge, S.M., et al., 2022. Heterogeneity in pragmatic randomised trials: sources and management. *BMC Medicine*, 20, 372. doi:10.1186/s12916-022-02569-w.

Hagemann, A., 2017. Cluster-robust bootstrap inference in quantile regression models. *Journal of the American Statistical Association*, 112, 446–456. doi:10.1080/01621459.2016.1148610.

Hoenig, J.M., and Heisey, D.M., 2001. The abuse of power. *The American Statistician*, 55 (1), 19–24. doi:10.1198/000313001300339897.

Hohberg, M., Pütz, P., and Kneib, T., 2020. Treatment effects beyond the mean using distributional regression: methods and guidance. *PloS One*, 15, e0226514. doi:10.1371/journal.pone.0226514.

Howard, M. 2018. *Comparison of the performance of simple linear regression and quantile regression with non-normal data: A simulation study*. Thesis (PhD). University of South Carolina.

Huber, M., and Wüthrich, K., 2019. Local average and quantile treatment effects under endogeneity: A review. *Journal of Econometric Methods*, 8 (1), 20170007. doi:10.1515/jem-2017-0007.

Kahan, B.C., and Morris, T.P., 2012. Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine*, 31, 328–340. doi:10.1002/sim.4431.

Koenker, R., 2005. *Quantile regression*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511754098.

Koenker, R., et al., 2017. *Handbook of quantile regression (Vol. 1st ed.)*. Chapman: Hall/CRC. doi:10.1201/9781315120256.

Koenker, R. 2021. *Quantreg: Quantile regression*. https://CRAN.R-project.org/package = quantreg.

Koenker, R., and Bassett-Jr, G., 1978. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46(1), 33–50.

Konstantopoulos, S., et al., 2019. Using quantile regression to estimate intervention effects beyond the mean. *Educational and Psychological Measurement*, 79 (5), 883–910. doi:10.1177/0013164419837321.

Kumle, L., Võ, M.L.H., and Draschkow, D., 2021. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, 53, 2528–2543. doi:10.3758/s13428-021-01546-0.

Lê Cook, B., and Manning, W.G., 2013. Thinking beyond the mean: A practical guide for using quantile regression methods for health services research. *Shanghai Archives of Psychiatry*, 25, 55–59. doi:10.3969/j.issn.1002-0829.2013.01.011.

O'Connell, N.S., … Gebregziabher, M., 2017. Methods for analysis of pre-post data in clinical research: A comparison of five common methods. *Journal of Biometrics & Biostatistics*, 08, 1–8. doi:10.4172/2155-6180.1000334.

Ohrnberger, J., et al., 2020. The worse the better? quantile treatment effects of a conditional cash transfer programme on mental health. *Health Policy and Planning*, 35, 1137–1149. doi:10.1093/heapol/czaa079.

Owen, K.L., et al., 2021. Implementation support improves outcomes of a fluency-based mathematics strategy: A cluster-randomized controlled trial. *Journal of Research on Educational Effectiveness*, 14 (3), 523–542. doi:10.1080/19345747.2021.1875526.

Parente, P.M.D.C., and Santos-Silva, J.M.C., 2016. Quantile regression with clustered data. *Journal of Econometric Methods*, 5 (1), 1–15. doi:10.1515/jem-2014-0011.

Petscher, Y., and Logan, J.A.R., 2014. Quantile regression in the study of developmental sciences. *Child Development*, 85 (3), 861–881. doi:10.1111/cdev.12190.

Tang, S., et al., 2021. A new quantile treatment effect model for studying smoking effect on birth weight during mother's pregnancy. *Journal of Management Science and Engineering*, 6 (3), 336–343. doi:10.1016/j.jmse.2021.06.005.
Tarr, G., 2012. Small sample performance of quantile regression confidence intervals. *Journal of Statistical Computation and Simulation*, 82 (1), 81–94. doi:10.1080/00949655.2010.527844.
Twisk, J., et al., 2018. Different ways to estimate treatment effects in randomised controlled trials. *Contemporary Clinical Trials Communications*, doi:10.1016/j.conctc.2018.03.008.
Tyler, E., et al., 2018. Research note: Collaborative institute for education research, evidence and impact (CIEREI). *Wales Journal of Education*, 20, 135–140. doi:10.16922/wje.20.1.8.
Yu, K., Lu, Z., and Stander, J., 2003. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52 (3), 331–350. doi:10.1111/1467-9884.00363.

Appendix 1: Plots indicating heterogeneity in residual variance.

MFaCTs: Grades 1-2 assessment

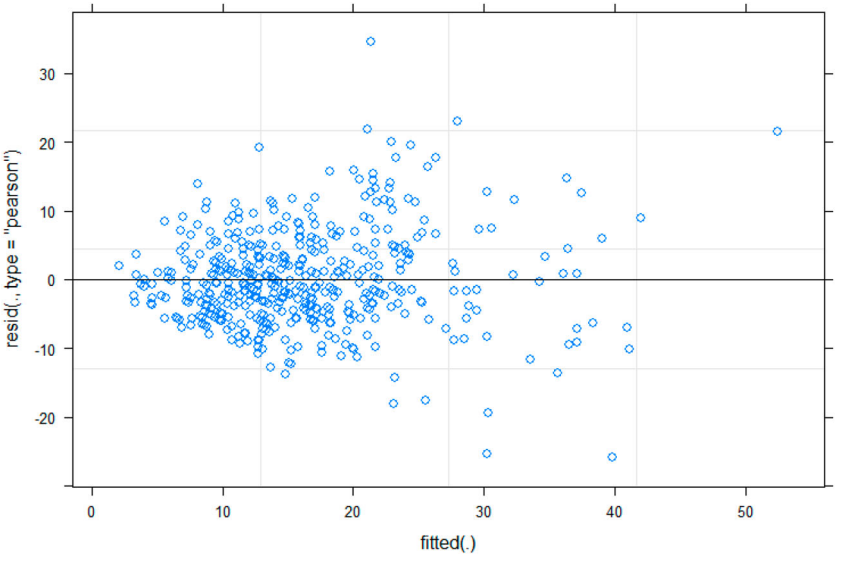

Figure A1 shows the Pearson residuals against fitted value. A clear funnel shaped pattern is present in the residual distribution indicating that residual variance is non-constant (heterogenious).

MFaCTs: Grades 3-4 assessment

Figure A2 shows the Pearson residuals against fitted value. Again a clear funnel shaped pattern is present in the residual distribution indicating that residual variance is non-constant (heterogenious).