25th International Congress on Modelling and Simulation, Darwin, NT, Australia, 9 to 14 July 2023
mssanz.org.au/modsim2023

A novel application of multilevel SEM: Teaching quality as mediator between intervention and student achievement

Ran Tian [a,b], Elizabeth Stojanovski [a] and Drew Miller [b]

[a] School of Information and Physical Sciences, The University of Newcastle, New South Wales
[b] Teachers and Teaching Research Centre, The University of Newcastle, New South Wales
Email: Ran.tian@newcastle.edu.au

Abstract: Improving teaching quality has been an ongoing pursuit for policymakers, researchers and broader government agencies and organisations. Professional development (PD) initiatives directed at teachers are a recognised strategy for improving teaching quality, with the underlying intention to support improved student achievement. Yet there is limited evidence of the actual impact of PD directed at teachers on student learning outcomes. Quality Teaching Rounds (QTR), a well-recognised form of collaborative PD in Australia, is one of few approaches to provide evidence of impact on teaching quality and student outcomes. Through multiple Randomized Controlled Trials (RCTs), a statistically significant positive impact of QTR PD has been demonstrated on teaching practices and student outcomes. A further question that remains is whether the change in teaching quality through involvement in QTR PD improves student achievement. This underlying mechanism is yet to be explored in the QTR PD framework and has rarely been explored in PD settings internationally due to limited data that often prevents suitable statistical techniques to be employed.

A clustered RCT investigating the impact of QTR PD on teaching quality and student achievement was conducted in 2019, involving 133 government primary schools in New South Wales (NSW), Australia. Data were collected from 222 teachers and 5146 students in Stage 2 (years 3-4) in Term 1 and these teachers and students were followed up in Term 4 of the same year. Teaching practices were observed and rated in the classroom using a pedagogical model, the Quality Teaching (QT) model, which contains 18 observable elements within three dimensions of teaching practice (Intellectual Quality, Quality Learning Environment and Significance), while students were assessed using the Progressive Achievement Test (PAT) in mathematics, which was measured using the scaled scores on 40 multiple choice questions. The data structure therefore comprised a combination of multilevel and longitudinal features along with latent constructs and multiple intervention groups (QTR vs wait-list control) that were being compared.

This paper examines the underlying interconnected relationships between PD, teaching quality and student achievement by testing the hypothesis that the impacts of the QTR intervention on student achievement in mathematics was mediated by teaching quality. Multilevel structural equation modelling (MSEM) with 2-2-1 design is investigated for these data. Student PAT scores in mathematics were significantly higher, on average (0.11SD [95% CI = 0.01,0.20]) in the intervention group (QTR) compared to those in the control group for the Intellectual Quality (IQ) dimension of the QT model. This demonstrates the statistically significant mediation effect of Intellectual Quality (IQ) on student learning outcomes.

Keywords: *Multilevel SEM, latent constructs, mediation, Monte Carlo confidence intervals, cluster RCT*

1. INTRODUCTION

The importance of education has been well recognized nationally and internationally as both the economic and non-economic benefits from investing in education are significant to both individuals and societies (Goczek et al., 2021; Hanushek et al., 2015; Hanushek and Woessmann, 2008; Patrinos and Psacharopoulos, 2013). Student achievement has been commonly used as a measure of the quality of education (Darling-Hammond, 2000). To assess improvements in the quality of an education system, various within- and between-school factors have been studied for their impacts on student learning. Teaching quality is considered a crucial within-school factor that influences student learning outcomes, with teaching quality usually measured by classroom observation. While Professional Development (PD) is a recognised key strategy for improving teaching quality (Darling-Hammond, 2000; Gore et al., 2017; Kennedy, 2006; Mizell, 2010), there is, to date, limited evidence of the effects of improved teaching quality on improved student achievement. Quality Teaching Rounds (QTR), a well-recognised PD approach in Australia, was designed as a pedagogy-based, collaborative approach to address limitations of traditional PD (Gore et al., 2017). The Quality Teaching (QT) model is one of two crucial components in QTR PD.

The QT model (NSW Department of Education and Training., 2003), a pedagogical classroom observation framework, is a revised version of the Productive Pedagogy (Ladwig, 2007) and Authentic Pedagogy (Newmann et al., 1996) models that preceded it. It was designed to assist school teachers to understand the underlying constructs of teaching and learning in New South Wales (NSW), Australia and can be applied across all learning stages (K-12) and Key Learning Areas (KLAs). Specifically, it contains a total of 18 elements across three dimensions, namely Intellectual Quality (IQ), Quality Learning Environment (QLE), and Significance (SIG), as shown in Table 1. Where IQ is about developing deep understanding of important knowledge; QLE is focused on ensuring positive classrooms that boost student learning; while SIG connects learning to students' lives and the wider community. Each dimension contains six observational elements, each element measured on a 5-point Likert scale. An element is defined as a component directly scored by observers/raters and a dimension as a grouping of elements that share a common feature. Latent constructs formed on the QT model is defined to measure teaching quality in the classroom and was formed for each of the three dimensions of the QT.

Table 1. The quality teaching model

Intellectual quality (IQ)	Quality learning environment (QLE)	Significance (SIG)
Deep knowledge	Explicit quality criteria	Background knowledge
Deep understanding	Engagement	Cultural knowledge
Problematic knowledge	High expectations	Knowledge integration
Higher order thinking	Social support	Inclusivity
Metalanguage	Students' self-regulation	Connectedness
Substantive communication	Student direction	Narrative

Despite the growing research on PD programs aimed to improve teaching practice and student achievement over the past two decades, large-scale RCT studies investigating both teaching quality and student achievement simultaneously are rare, mainly due to the costs and complexities involved with their implementation. This has led to a widely accepted, yet, rarely tested hypothesis that PD interventions influence the quality of teaching, which, in turn, influence student achievement. For the QTR PD, the impacts of QTR intervention on teaching quality and student academic achievement have been studied separately with results showing QTR PD to significantly affect teaching quality and student achievement in mathematics, when examined separately (Gore et al., 2017, 2021). This paves a solid foundation for further questioning the underlying mechanism between QTR PD, teaching quality and student achievement in mathematics. In this paper, we test the hypothesis that the effect of QTR PD intervention on student mathematics scores is mediated by teaching quality. Although cRCTs are commonly randomised at the school level, the measures of interest are at the individual level, resulting in a multilevel structure. The measures of primary interest for our study are at both teacher and student level, with students nested within teachers (i.e., a two-level structure). Additionally, our data includes latent constructs of the QT model and a longitudinal design with multiple intervention groups examined in the cRCT setting. Multilevel Structural Equation Modelling (MSEM) with a 2-2-1 design was thus employed. The methodology is presented in Section 2, followed by the results of modelling and associated discussion.

2. METHODOLOGY

2.1. Data

To investigate the impact of QTR PD on teaching quality and student achievement, a four-arm cluster Randomised Controlled Trial (cRCT) was conducted by the Teachers and Teaching Research Centre at the University of Newcastle in 2019. Described in detail in a protocol paper (Miller et al., 2019), this trial was arguably the largest cRCT conducted to investigate the impact of PD on both teaching quality and student achievement in Australia. Specifically, randomisation was at the school level using the Index of Community Socio-Educational Advantage (ICSEA). ICSEA is a measure of school socio-educational status which aims to make comparisons between schools more meaningful. Data was collected twice from the same group of Stage 2 (Year 3-4) teachers (with no prior QTR experience) and from Stage 2 students who were taught by these teachers, during school Term 1 (February - March) and again in Term 4 (October - November), 2019. Schools were randomly allocated to one of four intervention groups (including two QTR-related groups and two control-related groups) immediately after baseline data collection, with interventions undertaken in Term 2-3. The baseline data was to ensure groups to be similar enough for comparisons prior to interventions so that the differences between groups can be reasonably attributed to intervention effects (Sims and Fletcher-Wood, 2021). During baseline and follow-up data collection, up to two lessons were observed from each Stage 2 teacher and evaluated using the QT model by external research assistants across two school days while their students were assessed via an independent standardised Progressive Achievement Test (PAT) in mathematics that contained 40 multiple choice questions that were administered by research assistants to minimise potential biases.

The longitudinal cRCT data were collected from 133 government primary schools from NSW with 222 teachers and 5146 students, where 757 lessons were observed and rated in classrooms, and 10538 PATs were assessed. Based on preliminary analyses demonstrating no statistically significant difference between the two QTR-related groups and two control-related groups, the two QTR groups were consequently combined and the two control groups were also combined. Overall, the data structure for this study comprises a two-level structure, two intervention groups, a longitudinal design (baseline and follow-up) and three-factor latent constructs of the QT model.

2.2. Measurement model

Original constructs of the QT model shown in Table 1 were first examined by confirmatory factor analysis (CFA) using QTR RCT data. However, global fit indices including comparative fit index (CFI), the Tucker–Lewis index (TLI) and the Root Mean Square Error of Approximation (RMSEA) showed an unacceptable fit ($CFI<0.90$, $TLI<0.90$, $RMSEA>0.05$) to this data, potentially because the QT model was built for guiding teaching practice rather than for statistical analysis purposes. For this reason, explanatory factor analysis (EFA) was employed to obtain more objective latent constructs of the QT model to be able to statistically measure teaching quality. The 18 elements were separated into two sets (stable and unstable sets) following an EFA with different combinations of rotation methods, estimation methods and number of latent factors. Elements in the stable set remained consistently loaded onto the same latent factor across all combinations, in contrast to elements in the unstable set that loaded onto a range of different factors. The initial latent constructs formed using the stable elements were further examined using a CFA while unstable elements were moved across latent factors for an improved fit, as judged by three commonly used global fit indices (CFI, TLI, RMSEA) and local fit indices. The latent constructs formed by three pairwise correlated latent factors included IQ = {DK DU HE HOT SC EQC}, QLE = {E SS SSR} and SIG={BK PK KI C N}. Global fit indices of these latent constructs were satisfactory ($CFI>0.95$, $TLI>0.95$, $RMSEA<0.05$), with local fit indices also supporting well-fitting models. The reliability and validity of the constructs were tested were further confirmed by education experts to ensure that these latent constructs were practically meaningful.

2.3. Multilevel structural equation model

For multilevel data, the variables of interest are measured at different levels violating the independence assumption postulated in classic regression modelling. Consequently, estimates of standard errors in the classic regression setting are biased, which distorts statistical inference, including calculations of p-values and confidence intervals, which jeopardizes valid decision making. Linear mixed effects modelling (LMM) is, however, well suited to multilevel educational data to provide more accurate estimates of standard error and associated statistical inference. To test multilevel mediation, the LMM has been postulated as a suitable framework (see, for example, Krull and MacKinnon, 2001; Raudenbush and Bryk, 2002), which requires positive intraclass correlation coefficients (ICCs) and the outcome variable to be measured at the lowest level.

However, it suffers from major limitations including difficulties with handling latent constructs and multivariate outcomes, which implies that measurement error, commonly associated with educational data, is unable to be effectively incorporated within this modelling framework. Conflation of between- and within effects can occur, which leads to biased estimates at both levels (Lüdtke et al., 2008; Preacher et al., 2010). Additionally, the LMM framework cannot accommodate testing multilevel mediation effects in the presence of latent constructs. While structural equation modelling (SEM) can solve these limitations, these models cannot accommodate the multilevel data structure.

To assess school effectiveness, a multilevel structure co-exists (e.g., students nested within classes) with the situation that variables are often measured as latent constructs. Consequently, a model combining features from both LMM and SEM is required. Combining modelling from LMM into the SEM framework has been suggested since Schmidt's (1969) work. Several multilevel structural equation modelling (MSEM) methods, a generalised framework of LMM and SEM, have since been proposed, that enable models to accommodate the cluster/multilevel data structure and latent constructs within the single framework, allowing parameter estimation of all equations simultaneously, consequently improving efficiency and reducing bias. The MSEM framework is adopted from Muthen and Asparouhov (2008) for the present study, the general framework of which is depicted in Equations (1) – (3). This general framework was implemented in Mplus with the default estimation method, maximum likelihood estimation with robust standard errors (MLR), that do not require the standard normality assumption, with the between- and within effects separated by default, which avoids the conflation of between and within effects. The model is defined as follows.

Level-1 Measurement model: $$\boldsymbol{Y_{ij}} = \nu_j + \boldsymbol{\Lambda_j}\eta_{ij} + \epsilon_{ij} \quad (1)$$

Level-1 Structural model: $$\eta_{ij} = \alpha_j + \boldsymbol{B_j}\eta_{ij} + \boldsymbol{\Gamma_j X_{ij}} + \zeta_{ij} \quad (2)$$

Level-2 Structural model: $$\eta_j = \mu + \boldsymbol{\beta}\eta_j + \boldsymbol{\gamma X_j} + \zeta_j \quad (3)$$

where $\boldsymbol{Y_{ij}}$ is a $p \times 1$ vector of observed variables; ν_j is a vector of intercepts in the level-1 measurement model; $\boldsymbol{\Lambda_j}$ is a $p \times m$ matrix of factor loadings (weights); η_{ij} is a $m \times 1$ vector of latent variables; $\epsilon_{ij} \sim MVN(0, \Theta)$ is a vector of error terms. The level-1 measurement model serves to transform the observed variables to latent variables and can be used for dimension reduction if a latent construct exists. Otherwise, $\boldsymbol{\Lambda_j}$ reduces to an identity matrix and ν_j and ϵ_{ij} are ignored. This associated modelling procedure is referred to as path analysis. The α_j term is a vector of intercepts in the level-1 structural model; $\boldsymbol{B_j}$ is a $m \times m$ coefficients/correlations matrix among latent variables; $\boldsymbol{\Gamma_j}$ contains slope coefficients of covariates $\boldsymbol{X_{ij}}$ and $\zeta_{ij} \sim MVN(0, \Phi)$ is a vector of error terms. For the level-2 structural model, η_j contains all the elements of α_j (random intercepts) and $\boldsymbol{B_j}$ (random slopes if specified) that can vary at the level 2. $\boldsymbol{X_j}$ contains all level-2 covariates. μ contains the means of random effects and the intercepts of level-2 structural equations; $\boldsymbol{\beta}$ contains slope coefficients of random effects; $\boldsymbol{\gamma}$ contains slope coefficients of random effects regressed on level-2 covariates; $\zeta_j \sim MVN(0, \Psi)$ is a vector of error terms. Both structural models serve to simultaneously estimate multiple regression equations.

2.4. Multilevel mediation analysis in MSEM

MSEM for testing multilevel mediation is recommended, a general framework for which was proposed by Preacher et al., (2010). The single-level mediation model (Baron and Kenny 1986) has been shown to be a special case of the MSEM framework. As variables can be measured at different levels in a multilevel structure, a set of sub-models defined by levels that variables are measured at have been proposed for multilevel mediation analysis (Krull and MacKinnon, 2001; Preacher et al., 2010). Hereafter, we only consider the case where the independent variable, X, and mediators, M, were measured at level 2 and the dependent variable, Y, was measured at level 1, namely a 2-2-1 design. It is worth noting that mediation effects in the 2-2-1 design can only occur at the second/cluster level as there is no variation at the within level for the independent variable, X, or mediators, M. The general model depicted in Eq (1)-(3) can be reduced to better fit in the 2-2-1 design (see Appendix B in Preacher et al., 2010).

Regardless of the modelling framework, estimation and its inference of ab is central to mediation analysis. The most commonly used standard error of $\hat{a}\hat{b}$ was derived by the delta method and is commonly employed by the Sobel test (Sobel, 1982), or equivalently, for construction of a symmetric confidence interval (CI) of the indirect effect. It is a conservative approach and hence has reduced power to detect statistically significant indirect effects (MacKinnon et al., 1995). Furthermore, the product of two normal random variables is no longer normally distributed and the indirect effect tends to be skewed, which implies asymmetric CIs are more likely. Consequently, this method is neither theoretically nor practically optimal. Various methods can be

employed to construct asymmetric CIs for the indirect effect ab with the only difference being how the sampling distribution of $\hat{a}\hat{b}$ is obtained and used for inference (Preacher and Selig, 2012). A resampling bootstrap technique (Efron, 1979) is the most popular alternative to obtain asymmetric CIs, particularly for single-level mediated effects.

For a clustered data structure, bootstrap resamples the data in the same manner as the original data was sampled from the true population (Fox, 2005; Huang, 2018). However, implementation can be problematic if the data structure is mixed within clusters, has multiple intervention groups and has longitudinal features. Such data structure exists in our data. For the multilevel mediation analysis, the Monte Carlo CI (MCCI) (Preacher and Selig, 2012) is a popular alternative. This method shares most advantages of the bootstrap methods. Additionally, a unique advantage compared to bootstrap methods is that it produces results very fast, regardless of the number of replications and only needs to fit the model once. Additionally, performance of MCCIs is comparable to that of bootstrap methods (MacKinnon et al., 2004; Preacher and Selig, 2012). Overall, this is a competitive and potentially the only practically feasible method in situations where it is not easy to conduct the bootstrap.

To test the hypothesis that the impact of QTR intervention on student achievement in mathematics was mediated by teaching quality (measured by three-factor latent constructs), the MSEM 2-2-1 design enabled estimation of the 2-2 and 2-1 equations simultaneously with the built-in MLR estimation method that is robust to non-normality and non-independence. Data cleaning and MCCIs were performed in R (R Core Team, 2022) where the codes of MCCIs were adopted from Selig and Preacher (2008) and MSEM analysis was performed in Mplus Version 8.7 (Muthén and Muthén, 1998).

Comparisons of group means on latent factors were conducted. Structured means analysis (Aiken et al., 1994; Dimitrov, 2006) was employed and requires the assumption of measurement invariance (MI) across intervention groups at both baseline and follow-up. As a prerequisite for accurate comparisons of groups on a latent construct, MI assesses whether the latent construct has the same meaning for each group (control and intervention). Technically, MI includes four steps, namely configural invariance (the initial unconstrained model), weak/metric invariance (factor loadings restricted to be equal across groups), strong/scalar invariance (factor loadings and intercepts restricted to be equal across groups) and strict invariance (factor loadings, intercepts and residuals restricted to be equal across groups). It is practically sufficient for a construct to reach strong scalar invariance for comparing group means on latent factors.

3. RESULTS

Table 2 displays the results of MI tests. The difference of χ^2 for each comparison of adjacent models is statistically insignificant ($p>0.05$) indicating that a certain level of MI is met across intervention groups. With the assumption of strong invariance across intervention groups met at both baseline and follow-up, this ensures that comparisons of latent means across intervention groups can be made. Furthermore, this provides justification for restricting factor loadings and intercepts to be the same for the QTR and control groups at baseline and follow-up, respectively, in the MSEM.

Table 2. Results of the difference of χ^2 tests for MI across two intervention groups

	Model	χ2 (df)	$\Delta\chi$2(df)	P-value
Baseline	configural invariance	301.95 (148)		
	weak invariance	308.30 (159)	6.32 (11)	0.85
	strong invariance	323.58 (170)	15.28 (11)	0.17
Follow-up	configural invariance	241.64 (148)		
	weak invariance	252.30 (159)	10.66 (11)	0.47
	strong invariance	263.92 (170)	11.62 (11)	0.39

The global fit indices (RMSEA=0.01, CFI=0.96, TLI=0.96) indicate acceptable overall fit of this hypothesised model to the data, which is an essential precursor to examination of local fit, including estimates of parameters and standard errors. As shown in Figure 1, for the estimation of path a (i.e., the additional follow-up increase of teaching practice in QTR group compared to that in the control group), the difference of latent means at follow-up for each dimension was adjusted by its corresponding baseline latent construct. IQ was found to be statistically significantly greater by 0.50SD (SE=0.19, p=0.01) at follow-up in the QTR group compared to that in the control group; QLE increased by 0.21SD (SE=0.15, p=0.16) more at follow-up in the QTR group compared to that in the control group; and SIG statistically significantly increased by 0.57SD (SE=0.16, p=0.00) more at follow-up in the QTR group compared to that in the control group.

For the estimation of path b (i.e., the relationship between each dimension and teacher-level student mathematics scores at follow-up), student mathematics scores at baseline and follow-up were first separated into student- and teacher-level for bias reduction (Lüdtke et al., 2008). At the teacher level, student mathematics scores at follow-up were predicted by three pairwise correlated latent factors at follow-up, namely IQ_2, QLE_2 and SIG_2, after adjusting for baseline student mathematics scores. Student follow-up mathematics scores statistically significantly increased by 0.23SD (SE=0.11, p=0.04) in classrooms, on average, for a single unit increase in IQ at follow-up in the QTR group; Student follow-up mathematics scores decreased by 0.06SD (SE=0.09, p=0.49) at classrooms, on average, for a single unit increase in QLE at follow-up in the QTR group. Student follow-up mathematics scores decreased by 0.07SD (SE=0.08, p= 0.39) in classrooms, on average, for a single unit increase of SIG at follow-up in the QTR group. For the statistically insignificant paths for QLE and SIG, we believe these dimensions may not conceptually be reflected by the test scores.

As mentioned previously, the indirect effects $a_i b_i$ can only be examined at the same level, which was at the teacher level in our case. Student follow-up mathematics scores increased by 0.11 SD [95% MCCI: 0.01,0.20] more, on average, at the teacher level, mediated by IQ through QTR PD than that through business-as-usual PD in the schools sampled. However, QLE and SIG did not appear to have significant mediation effects, as judged by the respective 95% MCCIs.

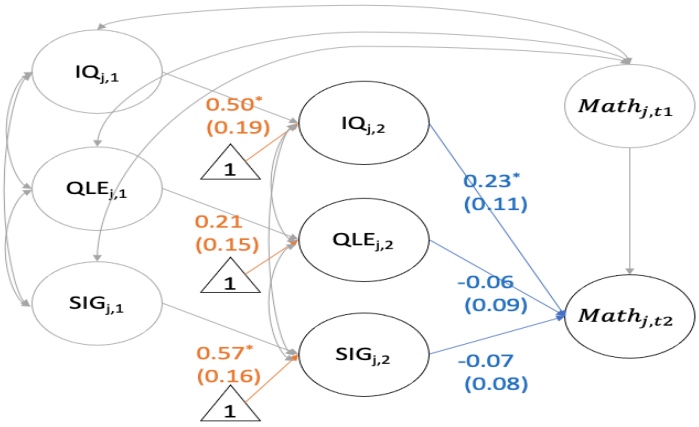

Figure 1. A simplified diagram of MSEM 2-2-1 design (teacher level only) with standardised estimates and standard errors for investigating the relationship between QTR intervention and student achievement in mathematics mediated by teaching quality (measured by three latent constructs). The orange paths indicate the additional increase of the corresponding dimension for the QTR group; the blue paths indicate the relationship between each factor and teacher-level student mathematics scores at follow-up. Global fit indices of the model are RMSEA=0.01, CFI=0.963, TLI=0.961, $*p < 0.05$. (Note: single arrows indicate the regression coefficients, circles indicate latent variables and double arrows indicate variance/covariance)

4. CONCLUSION

Multilevel mediation analysis in MSEM was investigated for this present study as a generalised framework that incorporates advantages of both LMM, which accommodates the multilevel data structure, and SEM that accommodates latent constructs. We found an additional 0.11 SD [95% MCCI = 0.01 0.20] of mediated effects attributed to teaching quality (through path IQ) at the intervention group on student mathematics scores in the intervention year. The only similar study we found was by Allen et al., (2011), which found an additional 0.06SD [95% CI = 0.01,0.13] mediated effects of teaching quality between their intervention and student achievement scores in the post-intervention year. Our study served two purposes: to further investigate the hidden mechanism in regards to how QTR PD, teaching quality and student achievement jointly work based on evidence from Gore et al., (2017, 2021); and to provide an example of dealing with real-world complex educational data using the MSEM framework.

REFERENCES

Aiken, L.S., Stein, J.A., Bentler, P.M., 1994. Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. Journal of Consulting and Clinical Psychology 62, 488–499. https://doi.org/10.1037/0022-006X.62.3.488

Allen, J.P., Pianta, R.C., Gregory, A., Mikami, A.Y., Lun, J., 2011. An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement. Science 333, 1034–1037.

Darling-Hammond, L., 2000. Teacher Quality and Student Achievement. Education Policy Analysis Archives 8, 1.

Dimitrov, D.M., 2006. Comparing groups on latent variables: A structural equation modeling approach 429–436.

Efron, B., 1979. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics 7, 1–26.

Fox, J., 2005. Applied Regression Analysis and Generalized Linear Models. SAGE Publications Inc.
Goczek, Ł., Witkowska, E., Witkowski, B., 2021. How Does Education Quality Affect Economic Growth? Sustainability 13, 6437. https://doi.org/10.3390/su13116437
Gore, J., Lloyd, A., Smith, M., Bowe, J., Ellis, H., Lubans, D., 2017. Effects of professional development on the quality of teaching: Results from a randomised controlled trial of Quality Teaching Rounds. Teaching and Teacher Education 68, 99–113. https://doi.org/10.1016/j.tate.2017.08.007
Gore, J.M., Miller, A., Fray, L., Harris, J., Prieto, E., 2021. Improving student achievement through professional development: Results from a randomised controlled trial of Quality Teaching Rounds. Teaching and Teacher Education 101, 103297. https://doi.org/10.1016/j.tate.2021.103297
Hanushek, E.A., Schwerdt, G., Wiederhold, S., Woessmann, L., 2015. Returns to skills around the world: Evidence from PIAAC. European Economic Review 73, 103–130. https://doi.org/10.1016/j.euroecorev.2014.10.006
Hanushek, E.A., Woessmann, L., 2008. The Role of Cognitive Skills in Economic Development. Journal of Economic Literature 46, 607–668. https://doi.org/10.1257/jel.46.3.607
Huang, F.L., 2018. Using Cluster Bootstrapping to Analyze Nested Data With a Few Clusters. Educ Psychol Meas 78, 297–318. https://doi.org/10.1177/0013164416678980
Kane, T.J., Staiger, D.O., 2012. Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project, Bill & Melinda Gates Foundation. Bill & Melinda Gates Foundation.
Kennedy, M.M., 2006. From Teacher Quality to Quality Teaching. Educational Leadership 63, 14–19.
Krull, J.L., MacKinnon, D.P., 2001. Multilevel Modeling of Individual and Group Level Mediated Effects. Multivariate Behav Res 36, 249–277. https://doi.org/10.1207/S15327906MBR3602_06
Ladwig, J.G., 2007. Modelling Pedagogy in Australian School Reform. Pedagogies: An International Journal 2, 57–76. https://doi.org/10.1080/15544800701343919
Lüdtke, O., Marsh, H.W., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., 2008. The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. Psychological Methods 13, 203–229. https://doi.org/10.1037/a0012869
MacKinnon, D.P., Lockwood, C.M., Williams, J., 2004. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. Multivariate Behav Res 39, 99.
MacKinnon, D.P., Warsi, G., Dwyer, J.H., 1995. A Simulation Study of Mediated Effect Measures. Multivariate Behav Res 30, 41. https://doi.org/10.1207/s15327906mbr3001_3
Miller, A., Gore, J., Wallington, C., Harris, J., Prieto-Rodriguez, E., Smith, M., 2019. Improving student outcomes through professional development: Protocol for a cluster randomised controlled trial of Quality Teaching Rounds. International Journal of Educational Research 98, 146–158.
Mizell, H., 2010. Why Professional Development Matters, Learning Forward (NJ). Learning Forward.
Muthen, B., Asparouhov, T., 2008. Growth mixture modeling: Analysis with non-Gaussian random effects 24.
Muthén, L.K., Muthén, B.O., 1998. Mplus User's Guide. Eighth Edition.
Newmann, F.M., Marks, H.M., Gamoran, A., 1996. Authentic Pedagogy and Student Performance. American Journal of Education 104, 280–312.
NSW Department of Education and Training., 2003. Quality teaching in NSW public schools: A classroom practice guide. Sydney, NSW: Department of Education and Training, Professional Support and Curriculum Directorate.
Patrinos, H.A., Psacharopoulos, G., 2013. Education: The Income and Equity Loss of not Having a Faster Rate of Human Capital Accumulation, in: Lomborg, B. (Ed.), How Much Have Global Problems Cost the World?: A Scorecard from 1900 to 2050. Cambridge University Press, Cambridge, pp. 170–191. https://doi.org/10.1017/CBO9781139225793.007
Preacher, K.J., Selig, J.P., 2012. Advantages of Monte Carlo Confidence Intervals for Indirect Effects. Communication Methods and Measures 6, 77–98. https://doi.org/10.1080/19312458.2012.679848
Preacher, K.J., Zyphur, M.J., Zhang, Z., 2010. A general multilevel SEM framework for assessing multilevel mediation. Psychological Methods 15, 209–233. https://doi.org/10.1037/a0020141
R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing.
Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods. Stage.
Schmidt, W.H., 1969. Covariance Structure Analysis of the Multivariate Random Effects Model (Ph.D.). The University of Chicago, United States -- Illinois.
Selig, J.P., Preacher, K.J., 2008. Monte Carlo method for assessing mediation: An interactive tool for creating confidence intervals for indirect effects.
Sims, S., Fletcher-Wood, H., 2021. Identifying the characteristics of effective teacher professional development: a critical review. School Effectiveness and School Improvement 32, 47–63.
Sobel, M.E., 1982. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. Sociological Methodology 13, 290. https://doi.org/10.2307/270723