OECD Education Working Papers No. 291

A Bayesian workflow for the analysis and reporting of international large-scale surveys: A case study using the OECD Teaching and Learning International Survey

David Kaplan,

Kjorte Harra

https://dx.doi.org/10.1787/588c4a12-en

Organisation for Economic Co-operation and Development

EDU/WKP(2023)6

For Official Use

English - Or. English

24 April 2023

DIRECTORATE FOR EDUCATION AND SKILLS

A Bayesian workflow for the analysis and reporting of international large-scale surveys: A case study using the OECD Teaching and Learning International Survey

OECD Education Working Paper No. 291

by David Kaplan and Kjorte Harra, University of Wisconsin - Madison

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

David Kaplan, david.kaplan@wisc.edu
Kjorte Harra, harra@wisc.edu
Gabor Fülöp, gabor.fulop@oecd.org
Marco Paccagnella, marco.paccagnella@oecd.org

JT03517528

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

OECD EDUCATION WORKING PAPERS SERIES

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

----

www.oecd.org/edu/workingpapers

Acknowledgements

This research was contracted by the Directorate for Education and Skills, Organisation for Economic Co-operation and Development (OECD). The authors are grateful to Gabor Fülöp, Aliénor Lerouge, Marco Paccagnella and Li Ruochen for their valuable comments on an earlier draft of this report. The authors also thank Emily Groves and Caio Passos Newman for editorial support. The comments and suggestions in this report are solely those of the authors and do not necessarily reflect the views of the OECD.

Abstract

This report aims to showcase the value of implementing a Bayesian framework to analyse and report results from international large-scale surveys and provide guidance to users who want to analyse the data using this approach. The motivation for this report stems from the recognition that Bayesian statistical inference is fast becoming a popular methodological framework for the analysis of educational data generally, and large-scale surveys more specifically. The report argues that Bayesian statistical methods can provide a more nuanced analysis of results of policy relevance compared to standard frequentist approaches commonly found in large-scale survey reports. The data utilised for this report comes from the Teaching and Learning International Survey (TALIS). The report provides steps in implementing a Bayesian analysis and proposes a workflow that can be applied not only to TALIS but to large-scale surveys in general. The report closes with a discussion of other Bayesian approaches to international large-scale survey data, in particular for predictive modelling.

Table of contents

Tables

Figures

1. Introduction

This report aims to showcase the value of implementing a Bayesian framework to analyse and report on data from international large-scale surveys with the Organisation for Economic Co-operation and Development (OECD) Teaching and Learning International Survey (TALIS) (OECD, 2019[1]; OECD, 2020[2]) serving as an example, and to provide guidance to users who want to analyse the data using this approach. The motivation for this report stems from the recognition that Bayesian statistical inference is fast becoming a popular methodological framework for the analysis of educational data generally, and large-scale surveys more specifically.

Bayesian inference can be conceptualised as a framework for quantifying uncertainty in statistical models. This uncertainty arises in not knowing (or ever knowing) the true value of a parameter of interest, for example a regression coefficient. This uncertainty is encoded into a Bayesian analysis through forming a probability distribution for the parameter(s) of interest describing the analyst's belief, before seeing the data, as to the central value and variance of a parameter. The analyst's prior beliefs can be more or less "informative", arising from a summary of past research, expert opinion, or both. The mechanics of Bayesian theorem (described in more detail below) combines prior beliefs with the extant data in hand to provide updated distributions of the parameters of interest. The major advantage of the Bayesian approach is how such results are interpreted. By explicitly assigning a probability distribution to parameter values, Bayesian analysis provides the framework to help answer questions such as "What is the most likely range of values for a given parameter?" or "What is the probability that a parameter exceeds a certain value?" The advantage of presenting results in this fashion is that it provides a much more nuanced analysis of the effects of interest, and is, arguably, much more informative to policy makers than simply indicating whether an effect is statistically significant or not.

1.1. Purpose and organisation of the report

The OECD published a two-volume report based on the results of TALIS 2018. The first volume was entitled *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners* (OECD, 2019[1]) and the second volume was entitled *TALIS 2018 Results (Volume II): Teachers and School Leaders as Valued Professionals* (OECD, 2020[2]). Both volumes not only contain detailed descriptive statistics across countries/economies, as well as by contextual variables, but, also, these volumes report the results of statistical models designed to provide predictive information regarding important outcomes of interest. For example, Volume II summarises the results of various regression analyses aimed at separately identifying relevant predictors of teacher job satisfaction and teacher self-efficacy – see Figures II.1.7 and II.1.8 in OECD (2020[2]). The analyses of these outcomes were carried out as follows: for each country, least-squares regression analysis was conducted with the TALIS scale of teacher job satisfaction or teacher self-efficacy as the dependent variable, and predictors, such as whether the teacher engaged in induction activities when joining the school. There were nine separate regression analyses. Many of these added control variables such as teachers' gender, years of experience as a teacher and self-efficacy. Sampling weights were also included, and sampling error was estimated using balanced repeated replication (BRR) weights to account for and adjust for the multi-stage, stratified, clustered nature of the sample. The results in Figures II.1.7 and II.1.8 of OECD (2020[2]) are displayed with marks indicating whether there was a positive and significant association (+) between job satisfaction and one of the predictors (after controls), a non-significant association with a blank mark, or a negative association (-) if there was a statistically significant negative association. The raw regression coefficients associated are also available in supplementary tables.

A major concern with the analytic approach used for the results in Figures II.1.7 and II.1.8 is that the categorisation of the results as positive, no effect, or negative, provides little information regarding the substantive importance of the effect in terms of how strongly different the results are from no effect at all. This issue touches on the continuing discussion over null hypothesis significance testing (Wasserstein and Lazar, 2016[3]) and the fact that, with large sample sizes such as those in TALIS, significant but relatively trivial results could be reported. Instead, it would be useful to have more substantive information regarding the importance of the effect, and the approach taken in this report is to compute the probability that the obtained effect is different from zero and to rank countries on the size of those probabilities. It is important to note that presenting results in this fashion can only be achieved via a Bayesian analysis, as will be described in more detail below. Therefore, the purpose of this report is to demonstrate an alternative mode of reporting based on reanalysing the Figures II.1.7 and II.1.8 from the TALIS report from the perspective of Bayesian statistical inference (Gelman et al., 2014[4]; Kaplan, forthcoming[5]).

The organisation of this report is as follows. In Section 2, we provide a brief overview of TALIS 2018. This is followed in Section 3 by a review of the key elements of Bayesian statistical inference that are relevant to this report. A more technical treatment of Bayesian inference is given in Kaplan (forthcoming[5]). In Section 4 we describe our analysis of the TALIS data as a special case of a so-called *Bayesian hierarchical model*, which incorporates the elements of multilevel modelling required for the proper analysis of data arising from complex sampling designs such as TALIS. Then, in Section 5 we present the steps of our reanalysis of Figures II.1.7 and II.1.8 in OECD (2020[2]). This will be followed in Sections 6 and 7 by the results of our reanalysis of teacher job satisfaction and teacher self-efficacy, respectively. We will display necessary diagnostic plots using data from the United States to demonstrate important aspects of Bayesian computation in Annex A. Also, we will provide both tables and figures of the estimates, as well as the probability of the obtained effects being different from zero, and then rank countries/economies by the sizes of these probabilities. We focus on only one analysis – namely the effect of participation in induction activities as it predicts teacher job satisfaction and teacher self-efficacy. Our reanalyses of the remaining predictors in Figures II.1.7 and II.1.8 are provided in Annex B and Annex C, respectively. Section 8 provides a proposed Bayesian workflow that can guide analyses of the type presented in this report, and Section 9 concludes with a discussion of the Bayesian advantage as it pertains to the analysis of international large-scale surveys (ILSA) data, as well as directions for future applications of Bayesian inference to ILSA data, particularly the problem of accounting for model uncertainty and prediction.

The R codes used for the analyses can be accessed from: https://gitlab.algobank.oecd.org/talisanalysis/wps/bayesian_workflow.

2. Overview of TALIS

In 2008, the OECD conducted the first cycle of TALIS. TALIS is an international, large-scale survey of teachers, school leaders and the learning environment in schools. The overarching goals of TALIS are to provide policy makers, educators, and other stakeholders with rigorous and detailed information around nine central themes. These have included: 1) teachers' instructional practices; 2) school leadership; 3) teachers' professional practices; 4) teacher education and initial preparation; 5) teacher feedback and development; 6) school climate; 7) job satisfaction; 8) teacher human resource issues and stakeholder relations; and 9) teacher self-efficacy.

2.1. Elements of the TALIS survey design

The first cycle of TALIS was conducted in 2008, the second cycle was conducted in 2013, and the third cycle on which this report is based was conducted in 2018. The fourth cycle will be conducted in 2024. Across the cycles of TALIS, the survey design has remained more or less unchanged. The key features of the TALIS design have focused on: 1) the identification of an international population of teachers and school leaders of mainstream schools, here defined as those teachers and school leaders working primarily in lower secondary (ISCED 2) schools; 2) a target sample size of 200 schools per country; 20 teachers and one school leader in each school; 3) a target response rate of 75% of the sampled schools, together with a 75% response rate from all sampled teachers in the country; 4) a target response rate of 75% of the sampled school leaders; 5) the construction of separate questionnaires for teachers and school leaders, each requiring between 45 and 60 minutes to complete; 6) two modes of data collection: questionnaires completed on paper or online, and 7) consistent survey windows for Northern and Southern Hemisphere countries.

2.2. TALIS reporting goals

As TALIS is a cross-sectional survey of teachers' and school leaders' attitudes, beliefs and opinions, it cannot be used to draw causal inferences. Instead, the strength of TALIS lies in its ability to provide internationally comparable evidence focused specifically on the day-to-day working lives of teachers and school leaders as seen from their perspective. This information is further broken down by relevant contextual variables such as teachers' gender, age and experience – and by schools' characteristics – geographical location, school type and composition. In addition, with information from the 2008 and 2013 cycles, important trend information can be gleaned to help inform country level policy. This is accomplished by keeping many of the survey questions constant across the cycles.[1]

[1] We argue later in this report that past cycles of TALIS, as well as the trend indicators, do not provide as much information as possible given new statistical methodologies.

3. Preliminaries on Bayesian inference

In this section, we provide a non-technical overview of Bayesian ideas. For a more technical review, see Gelman et al. (2014[4]) and Kaplan (forthcoming[5]).

Bayesian statistics has long been overlooked in the formal quantitative methods training of social scientists. Typically, the only introduction that a student might have had to Bayesian ideas is a brief overview of Bayes' theorem while studying probability in an introductory statistics class. This is not surprising. First, until recently, it was not feasible to conduct statistical modelling from a Bayesian perspective owing to its complexity and lack of available software. Second, Bayesian statistics addresses many of the problems associated with frequentist null hypothesis significance testing (Kaplan, forthcoming[5]; Wagenmakers, 2007[6]; Wasserstein and Lazar, 2016[3]), such as the methods applied to Figure II.1.7, and can, therefore, be controversial. We will use the term *frequentist* to describe the paradigm of statistics commonly used today, which also represents the counterpart to the Bayesian paradigm of statistics. Historically, however, Bayesian statistics predates frequentist statistics by about 150 years.

3.1. Frequentist probability

Following the discussion given in Kaplan (forthcoming[5]), most students and researchers in the social sciences were introduced to the axioms of probability by studying the properties of the coin toss or the dice roll. These studies address questions such as (1) What is the probability that the flip of a fair coin will return heads?; (2) What is the probability that the roll of two fair die will return a value of seven? To answer these questions requires enumerating the possible outcomes and then counting the number of times the event could occur. The probabilities of interest are obtained by dividing the number of times the event occurred by the number of possible outcomes - that is, the *relative frequency* of events. Before introducing Bayes' theorem, it is useful to review the axioms of probability that have formed the basis of frequentist statistics. These axioms of can be attributed primarily to the work of Kolmogorov (1956[7]).

Underlying frequentist statistics is the idea of the *long-run frequency*. An example of probability as long-run frequency concerns the dice roll. In this case, the number of possible outcomes of one roll of a fair die is six. If we wish to calculate the probability of rolling a two, then we simply obtain the ratio of the number of favourable outcomes (here there is only one favourable outcome), to the total possible number of outcomes (here six). Thus, the frequentist probability is 1/6 = 0.17. However, the frequentist probability of rolling a two is purely theoretical because, in practice, the die might not be truly fair, or the conditions of the toss might vary from trial to trial. Thus, the frequentist probability of 0.17, relates to the relative frequency of rolling a two in a very large (indeed infinite) and perfectly replicable number of dice rolls.

Nevertheless, this purely theoretical nature of long-run frequency plays a crucial role in frequentist statistical practice. Indeed, the entire structure of Neyman-Pearson hypothesis testing and Fisherian statistics that was used in the TALIS reports is based on the conception of probability as long-run frequency. Our conclusions regarding null and alternative hypotheses presuppose the idea that we could conduct the same study (in our case TALIS) an infinite number of times under perfectly reproducible conditions. Moreover, the frequentist interpretation of confidence intervals also assumes a fixed parameter with the confidence intervals varying over an infinitely large number of identical studies.

3.2. Epistemic probability

But there is another view of probability, and that is as *subjective belief*. Specifically, a modification of the Kolmogorov axioms was advanced by De Finetti (1974[8]; 1974[9]), who suggested replacing the (infinite) countable additivity axiom with finite additivity and also suggested treating probability as *subjective*.[2]

The use of the term *subjective* is perhaps unfortunate, insofar as it promotes the idea of fuzzy, unscientific, reasoning. Lindley (2007[10]) relates the same concern and prefers the term *personal probability* to *subjective probability*. Howson and Urbach (2006[11]) adopt the less controversial term *epistemic probability* to reflect an individual's greater or lesser degree of uncertainty about the problem at hand. Put another way, epistemic probability concerns *our uncertainty about unknowns*.

3.3. Bayesian inference

The goal of statistical inference is to obtain estimates of the unknown parameters, which we denote as θ. For this report, the unknown parameters will be regression coefficients relating policy-relevant predictors to key outcomes in TALIS. The major difference between Bayesian statistical inference and frequentist statistical inference concerns the assumptions regarding the nature of θ. In the frequentist tradition, the assumption is that θ is unknown, but has a fixed value that we wish to estimate. Measures such as the standard error or the frequentist confidence interval provide an assessment of the uncertainty associated with hypothetical repeated sampling from a population. In Bayesian statistical inference, θ is also considered unknown, however, similar to the data, θ is viewed as a random variable possessing a *prior probability distribution* that encodes our assumptions about the true value of θ before having seen the data. For example, on the basis of prior studies and/or expert opinion, we may be quite certain that the value of a regression coefficient of interest is positive, but uncertain about the range of values the coefficient can take on. In another case, we may also be quite certain about not only the sign of the effect but also its spread. Because both the observed data, denoted as y, and the parameters θ are assumed to be random variables, probability theory allows us to model the joint probability of the parameters and the data as a function of the conditional distribution of the data given the parameters, and the prior distribution, namely:

$$p(\theta, y) = p(y|\theta)p(\theta) \tag{1}$$

where $p(\theta, y)$ is the joint distribution of the parameters and the data, $p(y|\theta)$ is the distribution of the data conditional on the parameters and represents the expression of the model, and $p(\theta)$ is the prior distribution; again, the device wherein we encode our assumptions about the unknown parameters before seeing the data. Bayes' theorem (Bayes, 1763[12]; Laplace, 1951[13]) is then defined as:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{2}$$

Where $p(\theta|y)$ is referred to as the *posterior distribution* of the parameters θ given the observed data y representing our updated knowledge about the parameters of interest after having encountered the model and the data and is equal to the data distribution $p(y|\theta)$

[2] A much more detailed set of axioms for subjective probability was advanced by Savage (1954[37]).

times the prior distribution of the parameters $p(\theta)$ normalised by $p(y)$ so that the posterior distribution sums (or integrates) to one.

3.4. Prior distributions

The general approach to considering the choice of a prior distribution on θ is based on how much information we believe we have *prior* to data collection and how precise we believe that information to be. The strength of Bayesian inference lies in its ability to incorporate our uncertainty about θ directly into our statistical models.

3.4.1. Non-informative priors

In some cases, we may not be in possession of enough prior information to aid in drawing posterior inferences. Or, from a policy perspective, it may be prudent to not reveal assumptions about effects of interest and, instead, let the data speak for itself. Regardless, from a Bayesian perspective, this real or assumed lack of information is still important to consider and incorporate into our statistical models (Kaplan, forthcoming[5]).

The standard approach to quantifying a lack of information is to incorporate non-informative prior distributions into our analyses. In the case in which there is no prior knowledge to draw from, perhaps the most extreme non-informative prior distribution that can be used is the *uniform distribution* ranging from $-\infty$ to $+\infty$, and denoted as $U(-\infty, +\infty)$. The uniform distribution essentially signals that we believe that our parameter of interest can take on an infinite number of values, each of which is equally likely. The problem with this particular specification of the uniform prior is that it is not proper insofar as the distribution does not integrate to 1. However, this does not always lead to problems, and is more of a conceptual issue. Highly diffused priors such as the Gaussian distribution with a mean of zero and variance of ten, denoted as $\mathcal{N}(0,10)$, could also be used.

3.4.2. Weakly informative priors

Situated between non-informative and informative priors are *weakly informative* priors. Weakly informative priors are distributions that provide one with a method for incorporating less information than one actually has in a particular situation. Specifying weakly informative priors can be useful for many reasons. First, it is doubtful that one has complete ignorance of a problem for which a non-informative prior, such as the uniform distribution, is appropriate. Rather, it is likely that one can consider a more reasonable bound on the uniform prior, but without committing to much more information about the parameter. Second, weakly informative priors are very useful in stabilising the estimates of a model, particularly in cases of small sample sizes (Gelman, 2006[14]). Specifically, Bayesian inference can be computationally demanding and so, although one may have information about, say, higher level variance terms, such terms may not be substantively important, and/or they may be difficult to estimate, especially in small samples. Therefore, providing weakly informative prior information may help stabilise the analysis without impacting inferences.

3.4.3. Informative priors

Finally, it may be the case that, on the basis of previous research, expert opinion, or both, information can be brought to bear on a problem and be systematically incorporated into the prior distribution. Such priors are referred to as *informative*. Informative prior distributions require that the analyst commit to the shape of the distribution. For example, if a parameter of interest, such as a regression coefficient, is assumed to have a normal prior distribution, then the analyst must commit to specifying the average value and the precision around that value. Given that informative priors are inherently subjective in nature, they can be quite incorrect. Fortunately, Bayesian theory provides numerous methods for assessing the sensitivity of results to the choice of prior distributions.

3.5. Bayesian computation in brief

As stated in the introduction, the key reason for the increased popularity of Bayesian methods in the social sciences has been the (re)discovery of numerical algorithms for estimating posterior distributions of the model parameters given the data. Prior to these developments, it was virtually impossible to derive summary measures of the posterior distribution, particularly for complex models with many parameters. The numerical algorithms that we will describe in this chapter involve Monte Carlo integration using Markov chains - also referred to as *Markov chain Monte Carlo* (MCMC) sampling. These algorithms have a rather long history, arising out of statistical physics and image analysis (Geman and Geman, 1984[15]; Metropolis N. et al., 1953[16]). For a nice introduction to the history of MCMC see Robert and Casella (2011[17]).

Bayesian inference focuses on calculating summary statistics of the posterior distribution. For very simple problems, this can be handled analytically. However, for complex, high-dimensional problems involving multiple integrals, the task of analytically obtaining summary statistics can be virtually impossible. So, rather than attempting to analytically solve these high dimensional problems, we can instead use well-established mathematical computation methods to draw samples from a *target distribution* of interest (in our case the posterior distribution) and summarise the distribution formed by those samples. This is referred to as *Monte Carlo integration.*

Often, we direct the algorithm to sample from multiple points in the posterior distribution. These are referred to as *chains*, and our goal is to ensure that the MCMC samples arising from each chain *mix* well and yield a good approximation to the true posterior distribution of each of the model parameters. In addition, the nature of MCMC algorithms is to initiate dependent draws from the posterior distribution with the goal that over the iterations, the draws become independent. This is important for monitoring the so-called *effective sample size* of the analysis. Strong autocorrelation over the iterations yields draws that are not independent and hence lead to lower effective sample sizes on which the posterior estimates are obtained. The converse is that lower autocorrelation indicates independent draws and effective sample sizes that are close to the actual number of draws requested of the algorithm. An approach to aiding in reducing autocorrelation is to calculate posterior statistics based on every t^{th} draw from the posterior distribution. This is called *thinning*.

Given the computational complexity of MCMC, it is absolutely essential for Bayesian inference that the convergence of the MCMC algorithm be assessed. The importance of assessing convergence stems from the very nature of MCMC in that it is designed to converge to a distribution rather than to a point estimate. Because there is not a single adequate assessment of convergence, it is important to inspect a variety of diagnostics that examine varying aspects of convergence. Among these are: (a) trace plots for mixing, (b) autocorrelation plots to assess independence, (c) posterior probability distribution (density) plots for all parameters to assess mixing and convergence, (d) potential scale reduction factors to assess mixing and convergence, and (e) effective sample size to assess independence. Of these, we will concentrate primarily on the potential scale reduction factor (referred to as $Rhat$), and the effective sample size (referred to as *n_eff*) as these two provide the most reliable information regarding the convergence of the algorithm. These diagnostics will be used in this report.

3.6. Summarising the posterior distribution

Having obtained satisfactory convergence to the posterior distribution, the next step is to calculate point estimates and obtain relevant intervals. The expressions for point estimates and intervals of the posterior distribution come from expressions of conditional distributions generally.

3.6.1. Posterior predictive checking

A very natural way of evaluating the overall quality of a model is to examine how well the model fits the actual data. Examples of such approaches abound in frequentist statistics, often based on "badness-of-fit" measures. In the context of Bayesian statistics, the approach to examining how well a model fits the data is based on the notion of *posterior predictive checking*, and the accompanying *posterior predictive p-value*. An important philosophical defence of the use of posterior predictive checks can be found in Gelman and Shalizi (2012[18]).

The general idea behind posterior predictive checking is that there should be little, if any, discrepancy between data generated by the model, and the actual data itself. Any deviation between the data generated from the model and the actual data implies model misspecification.

In the Bayesian context, the approach to examining model fit and specification utilises the posterior predictive distribution of replicated data accounting for uncertainty via the priors that are placed on the model parameters. Thus, posterior predictive checking accounts for the uncertainty in the model parameters and the uncertainty in the data.

As a means of assessing the fit of the model, posterior predictive checking implies that the replicated data should match the observed data quite closely if we are to conclude that the model fits the data. One approach to quantifying model fit in the context of posterior predictive checking is to calculate the posterior predictive *p*-value. If the model-generated data fit the actual data well, then any differences should be due to chance – meaning that the posterior *p*-value should be around 0.50. Any large deviations suggest model misfit that could stem from model misspecification (e.g. omitted variables, incorrect functional form), poorly specified priors, or both.

3.6.2. Interval summaries of the posterior distribution

One important consequence of viewing parameters probabilistically concerns the interpretation of *uncertainty intervals*. Recall that the frequentist confidence interval requires that we imagine a fixed parameter, say the population mean μ. Then, we imagine an infinite number of repeated samples from the population characterised by μ. For any given sample, we can obtain the sample mean $\bar{x}$ and then form a $100(1-\alpha)\%$ confidence interval. The correct frequentist interpretation is that $100(1-\alpha)\%$ of the confidence intervals formed this way capture the true parameter μ under the null hypothesis. Notice that, from this perspective, the probability that the parameter is in the interval is either zero or one.

In contrast, the Bayesian framework assumes that a parameter has a probability distribution. Sampling from the posterior distribution of the model parameters, we can obtain its quantiles. From the quantiles, we can directly obtain the probability that a parameter lies within a particular interval. So, for example, a 95% posterior probability interval (also referred to as a *credible interval*) would mean that the probability that the true value of the parameter lies in the interval is 0.95. Notice that this is entirely different from the frequentist interpretation, and arguably aligns with common sense.[3]

Symmetric intervals such as the 95% posterior probability interval are not the only interval summaries that can be obtained from the posterior distribution, and a major benefit of Bayesian inference is that any interval of substantive importance can be obtained directly from the posterior distribution through simple functions available in R. This is particularly noteworthy when trying to gauge just how much different an obtained estimated effect is from zero. That is, even if zero lies within the 95% credible interval, there may be a sizable difference between zero and the obtained effect in terms of the distribution of credible values. We present these probabilities in this report, but it should further be noted that the flexibility available in being able to summarise any aspect of the posterior distribution admits a much greater degree of nuance in the kinds of research questions one may ask.

[3] Interestingly, the Bayesian interpretation is often the one incorrectly ascribed to the frequentist interpretation of the confidence interval.

4. Analysis of TALIS data as a Bayesian hierarchical model

A common feature of data collection in the social sciences is that units of analysis (e.g. students or employees) are nested in higher level organisational units (e.g. schools or companies, respectively). Indeed, in many instances, the substantive problem specifically concerns an understanding of the role that units at both levels play in explaining or predicting outcomes of interest. For example, the TALIS study deliberately samples schools (within a country) and then samples teachers within the sampled schools. Such data collection plans are generically referred to as *clustered sampling designs*. Data from clustered sampling designs are then collected at both levels for the purpose of understanding each level separately, but also to understand the inputs and processes of teacher and school level variables as they predict both school and teacher level outcomes. Higher levels of nesting are, of course, possible: e.g. teachers nested in schools, which, in turn, are nested in local educational authorities, such as school districts.

It is probably without exaggeration to say that one of the most important contributions to the empirical analysis of data arising from such data collection efforts has been the development of so-called *multilevel models*. Original contributions to the theory of multilevel modelling for the social sciences can be found in Burstein (1980[19]), Goldstein (2011[20]), and Raudenbush and Bryk (2002[21]), among others.

4.1. The intercepts and slopes as outcomes model

For this report, we discuss the most general form of the multilevel model - the intercepts and slopes as outcomes model, with an example that will be presented below. Suppose that interest centres on reported job satisfaction of teachers in the United States. Denote $t3jobsa_{ij}$ as reported job satisfaction of teacher i in school j. We may wish to model $t3jobsa_{ij}$ as a function of whether teacher i took part in induction activities in school j, denoted as $tt3g08_{ij}$. The intercepts and slopes as outcomes model can be written as:

$$t3jobsa_{ij} = \beta_{0j} + \beta_{1j}\,(tt3g08)_{ij} + r_{ij} \tag{3}$$

where β_{0j} is the intercept for school j, representing the average job satisfaction score for the school, β_{1j} are the regression coefficients representing the relationship between teacher job satisfaction and career choice, which might vary over the J schools, and r_{ij} is a residual term. Raudenbush and Bryk (2002[21]) have referred to the model in Equation (3) as the "level-1" model.

Interest in multilevel regression models stems from the fact that we can model the intercepts and slopes as a function of school level predictors, which we will denote as $\boldsymbol{z}_j$. For example, we could ask whether school average job satisfaction or the relationship between school average job satisfaction and first career choice can be predicted by whether the school is public or private. For the TALIS reports that we are reanalysing, school level effects were not included but, rather, the intercepts and slopes were allowed to simply vary across schools without an attempt to explain the variation. In this case, the so-called *level-2* model can be written as:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{4a}$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \tag{4b}$$

where γ_{00} is the grand mean of job satisfaction, γ_{10} the grand mean of the job satisfaction and induction relationship, and u_{0j} and u_{1j} capture unmodeled between-school variation.

To express Equations (3) and (4a) - (4b) as a Bayesian hierarchical model, we specify the following distributions for $t3jobsa_{ij}$, β_{0j}, β_{1j}, γ_{00}, and γ_{01}. That is,

$$t3jobsa_{ij} \sim \mathcal{N}\left[\beta_{0j} + \beta_{1j}(tt3g08)_{ig}, \sigma_g^2\right] \quad (5a)$$

$$\beta_{0j} \sim \mathcal{N}(\gamma_{00}, \tau_{00}^2) \quad (5b)$$

$$\beta_{1j} \sim \mathcal{N}(\gamma_{10}, \tau_{10}^2) \quad (5c)$$

$$\gamma_{00} \sim \mathcal{N}(\mu_{00}, \omega_{00}^2) \quad (5d)$$

$$\gamma_{10} \sim \mathcal{N}(\mu_{10}, \omega_{10}^2) \quad (5e)$$

To complete the hierarchical specification, prior distributions would need to be supplied for the variance terms, σ_j^2, τ_{00}^2, τ_{10}^2, ω_{00}^2, and ω_{10}^2. Several reasonable choices of priors are available for the variance terms but, for this report, we chose a non-informative half-Cauchy distribution because it has been shown to be computational stable as a non-informative prior for variance terms (Gelman, 2006[14]; Kaplan, forthcoming[5]).

5. An example using TALIS

The following section discusses the specifics of how analyses were conducted for this report. We begin by describing the TALIS sample and then move on to how we treat missing data and sampling weights.

5.1. Sample

The data used in these analyses originates from the 2018 cycle of TALIS, which includes 48 countries and economies. TALIS focuses on teachers and school leaders in lower secondary education. TALIS follows a stratified two-stage probability sampling design. This means that teachers are randomly selected from the list of in-scope teachers for each of the randomly selected schools.

5.2. Missing data

Missing responses coded as "not reached" or responses that were otherwise omitted or deemed invalid were imputed using *predictive mean matching* (Rubin, 1986[22]). The essential idea behind predictive mean matching is that missing values are imputed by matching the predicted values from the observed data using a predictive mean metric to the predicted values using regression imputation. Then, the procedure uses the actual observed value for the imputation. That is, for each regression, there is a predicted value for the missing data as well as a predicted value for the observed data. The predicted value for the observed data is then matched to a predicted value of the missing data using, say, a nearest neighbour distance metric. Once the match is found, the actual observed value (rather than the predicted value) replaces the missing value. If more than one match is found, a random match is used.

Although this process can be conducted only once to impute missing data, multiple draws of plausible values account for uncertainty surrounding a single imputed missing data point. The practice of multiple imputation fits in the Bayesian perspective, as parameters are assumed to take on a probability distribution instead of a singular fixed, but unknown, value. For the current analysis, ten predictive mean matching imputations were computed using the *mice* package in *R* (van Buuren, 2012[23]).

Due to the format of the TALIS teacher survey, several questions were not logically applicable to a given respondent due to their answers to previous questions. These missing patterns are not as easily imputable as the seemingly random missing responses previously discussed. For example, a teacher who indicated they never participated in any induction activities at their current school would not answer further questions asking for details about what kinds of induction activities they participated in. Here, missing responses for the following questions specifying their participation in activities were imputed to show they did not participate in that specific activity. For other questions where there was not a logically imputed response from not-applicable questions or questions that were not administered in certain countries, responses were excluded from the analysis.

5.3. Sampling weights

The use of weights is important because it allows researchers to conduct statistical analyses using non-representative samples that can be mathematically corrected to better represent the population of interest. With a survey such as TALIS, schools are randomly sampled in a given country or economy, and teachers within those schools are sampled. These samples cannot be perfectly representative of the population of teachers in a given country, so the implementation of weights to counteract this is necessary.

For this study, we use the final estimation weights, denoted as $TCHWGT_{hij}$, which were drawn from the original TALIS 2018 report. These weights are calculated as the product of design weights for schools, the design weight for teachers, and the three adjustment factors for teachers. However, preliminary analyses revealed that more stable convergence of the computing algorithm could be achieved through the use of normalised sampling weights. These normalised weights, denoted as $NormWgt_{hij}$, were calculated for each participating teacher as the ratio of sample size n to the total population N multiplied by the final estimation weights $TCHWGT_{hij}$. The normalised weight can be written as:

$$NormWgt_{hij} = \frac{n}{N} * TCHWGT_{hij} \qquad (6)$$

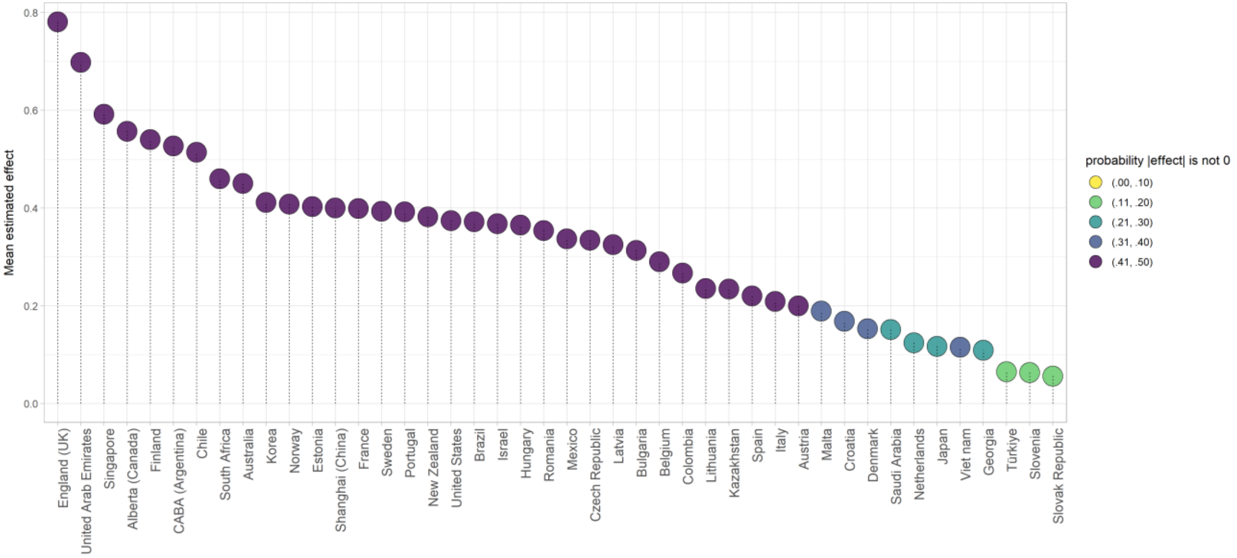
where i denotes each participating teacher for each participating school j in explicit stratum h for sample size n and total population size of N.

6. Results for the analysis of teachers' job satisfaction

All analyses used the *Stan*-based software programme *rstanarm* (Goodrich et al., 2020[24]; Stan Development Team, 2021[25]; Stan Development Team, 2021[26]). We requested four chains with 5 000 iterations per chain. The algorithm uses half of the iterations as warm-up, and we requested a thinning interval of 10. This leads to a total sample size of 1 000 iterations. Annex A the convergence plots for the analysis of teacher job satisfaction and teacher self-efficacy from the United States sample only. The plots suggest some small concerns regarding convergence, but the *Rhat* and n_eff values (Table 1) reveal adequate evidence of convergence. Despite this, the analysis shows a relatively poor fit to country-average teacher job satisfaction with a posterior predictive p-value of 0.67 (Figure A A.4). Similar convergence plots and posterior predictive checks would be necessary for each country and for all analyses.

Figure 1 displays the results of the regression of teacher job satisfaction on participation in any induction activities at a teacher's current school, controlling for a teacher's gender and years of experience as a teacher, ordered in terms of the size of the effect, labelled on the y-axis. The colour of the bubble represents the probability that the effect (i.e. the mean of the posterior distribution) is different from zero. We believe this plot (and subsequent plots and tables) conveys the idea that an effect could be deemed non-significant from a frequentist point of view (i.e. not significantly different from zero) but that the actual difference between the obtained effect and zero could be quite large. Note that, because the estimated effect is at the mean of the posterior distribution, these probabilities cannot exceed 0.5 in absolute value.

Figure 1. Participation in any induction activities at current school and teachers' job satisfaction



Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

We see in Figure 1 that countries such as England (United Kingdom) and the United Arab Emirates showed larger mean estimated effects and had larger probabilities that the effect is different than zero. This is perhaps not surprising; however, note that there are countries with smaller effects but which, nevertheless, have similarly large probabilities of the effect being greater than zero, such as Austria and Italy. We also observe more uncertainty in parameter estimates for countries with smaller estimated effects, such as in Slovenia and Türkiye. Bubble plots are provided for all of the regression analyses included in Figure II.1.7 of OECD (2020[2]) and can be found in Annex B.

These bubble plots are designed to provide a quick glance at the results. More detailed results, including the 95% posterior probability interval and the precise probability that the effect of interest is different than zero can be found in Table 1, where we also present the results of the least-squares regression from OECD (2020[2]).

Table 1. Participation in any induction activities predicting teachers' job satisfaction

Country/economy	Effective sample size	Rhat	Posterior mean (sd)	95% CI		$P(\|effect\|) \neq 0$	Original results[1]
				2.5%	97.5%		
England (UK)	763	1	0.78 (0.21)	0.37	1.21	0.50	**0.84**
United Arab Emirates	1 081	1	0.70 (0.10)	0.51	0.88	0.50	**0.65**
Singapore	917	1	0.59 (0.16)	0.27	0.92	0.50	**0.63**
Finland	947	1	0.54 (0.16)	0.23	0.88	0.50	0.48
CABA (Argentina)[2]	955	1	0.53 (0.15)	0.23	0.82	0.50	**0.54**
Chile	943	1	0.51 (0.19)	0.13	0.89	0.50	**0.44**
Portugal	930	1	0.49 (0.13)	0.13	0.64	0.50	**0.44**
Australia	935	1	0.45 (0.16)	0.14	0.76	0.50	0.52
Korea	1 018	1	0.41 (0.16)	0.10	0.71	0.50	0.38
Norway	1 016	1	0.41 (0.11)	0.17	0.62	0.50	**0.39**
Estonia	971	1	0.40 (0.12)	0.17	0.64	0.50	**0.38**
Shanghai (China)	1 031	1	0.40 (0.12)	0.17	0.63	0.50	**0.43**
Brazil	1 072	1	0.37 (0.13)	0.12	0.65	0.50	**0.46**
Hungary	1 015	1	0.37 (0.14)	0.10	0.63	0.50	**0.34**
Romania	1 004	1	0.35 (0.11)	0.14	0.56	0.50	**0.39**
Mexico	1 171	1	0.34 (0.10)	0.15	0.53	0.50	**0.29**
Czech Republic	954	1	0.33 (0.11)	0.14	0.54	0.50	**0.36**
Latvia	997	1	0.33 (0.12)	0.08	0.56	0.50	0.31
Belgium	936	1	0.29 (0.10)	0.09	0.49	0.50	0.37
Kazakhstan	1 056	1	0.23 (0.07)	0.10	0.36	0.50	**0.27**
Alberta (Canada)	1 053	1	0.56 (0.24)	0.10	1.03	0.49	0.53
South Africa	762	1	0.46 (0.19)	0.08	0.85	0.49	**0.79**
France	607	1	0.40 (0.17)	0.06	0.73	0.49	**0.43**
Sweden	712	1	0.39 (0.18)	0.03	0.73	0.49	**0.38**
Colombia	1 074	1	0.27 (0.12)	0.05	0.51	0.49	**0.26**
Spain	960	1	0.22 (0.09)	0.04	0.40	0.49	**0.31**
New Zealand	1 010	1	0.38 (0.18)	0.02	0.73	0.48	**0.34**
United States	1 048	1	0.37 (0.18)	0.03	0.72	0.48	**0.34**
Israel	1 056	1	0.37 (0.17)	0.04	0.69	0.48	**0.27**
Bulgaria	1 021	1	0.31 (0.15)	0.03	0.60	0.48	**0.25**
Austria	1 101	1	0.20 (0.12)	-0.03	0.45	0.45	0.21
Italy	1 101	1	0.21 (0.13)	-0.07	0.69	0.44	**0.27**
Croatia	1 016	1	0.17 (0.13)	-0.08	0.43	0.40	**0.21**
Viet Nam	932	1	0.12 (0.10)	-0.08	0.31	0.37	**0.09**
Malta	984	1	0.19 (0.21)	-1.90	0.64	0.32	0.28
Denmark	1 020	1	0.15 (0.17)	-0.20	0.52	0.31	**0.27**

Country/economy	Effective sample size	Rhat	Posterior mean (sd)	95% CI		$P(\|effect\|) \neq 0$	Original results[1]
				2.5%	97.5%		
Netherlands	1 073	1	0.12 (0,15)	-0.17	0.32	0.29	**0.13**
Saudi Arabia	986	1	0.15 (0.20)	-0.23	0.54	0.28	**0.40**
Japan	903	1	0.12 (0.15)	-0.17	0.42	0.28	0.13
Georgia	990	1	0.11 (0.18)	-0.22	0.46	0.23	**0.08**
Slovak Republic	963	1	0.06 (0.11)	-0.16	0.27	0.19	**0.10**
Slovenia	1 055	1	0.06 (0.14)	-0.20	0.33	0.18	**0.04**
Lithuania	1 094	1	0.24 (0.10)	0.04	0.44	0.16	**0.26**
Türkiye	1 030	1	0.07 (0.16)	-0.25	0.38	0.15	0.13

Notes:
1. Statistically significant results from OECD (2020[2]) are reported in bold.
2. CABA (Argentina): Ciudad Autónoma de Buenos Aires, Argentina.
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

An inspection of Table 1 reveals that, across the countries and economies, the effective sample size is nearly 1 000, indicating low autocorrelation. In addition, the *Rhat* values are 1.0, indicating convergence of the algorithm. Further inspection of Table 1 provides insight into one of the main advantages of using Bayesian methods for the analysis and reporting of ILSA data, namely the capacity to examine the entire posterior distribution of the effect. Take, for example, Austria and Georgia. For Austria, we observe that zero is in the credible interval and its frequentist *p*-value also indicates that the effect is not statistically significant. Yet, the probability that the effect is greater than zero is 0.45. So, the *p*-value (and the frequentist confidence interval) would lead to a single decision of non-significance, and the credible interval would indicate that the zero is a plausible value. However, because we have the whole posterior distribution to work with, the actual probability that the effect is greater than zero is 0.45, which is arguably large. Contrast this with Georgia, where zero is in the credible interval but the effect is deemed statistically significant. However, the actual probability that the effect is greater than zero is small (0.22).

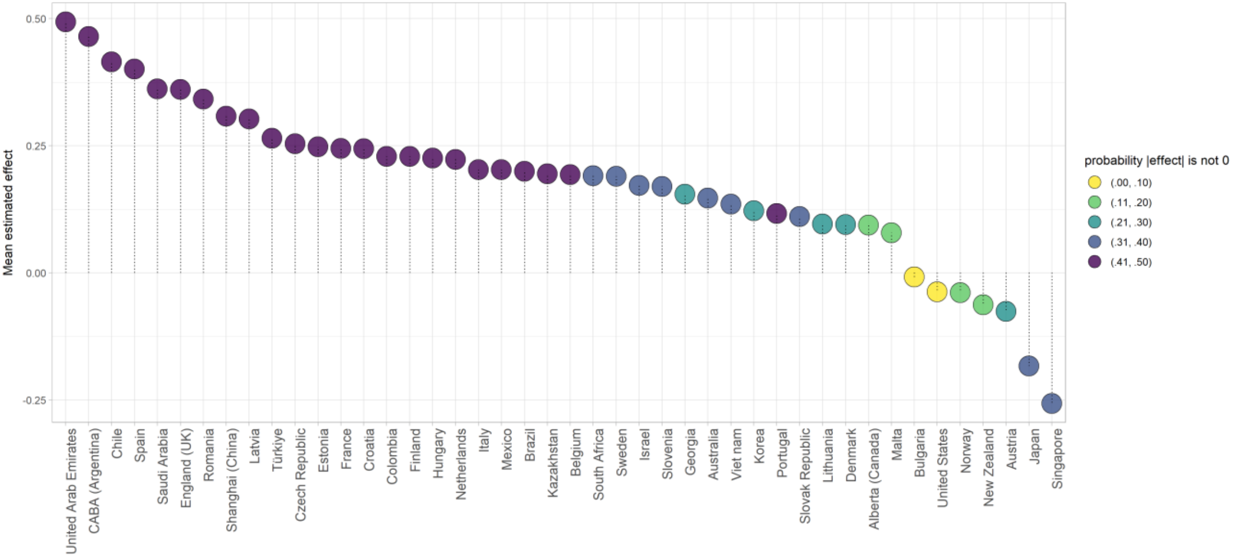
Of course, these interpretations require substantive justification, but we argue that the presentation of posterior effect probabilities provide more nuance, and arguably more policy-relevant information for cross-country comparisons, while accounting for more uncertainty than conventional significance tests. It must be emphasised that this type of interpretation is only possible because of access to the entire posterior distribution of the effect – a consequence of Bayesian inference – and not possible via conventional significance testing.

7. Results for the analysis of teachers' self-efficacy

Annex A also displays the convergence diagnostic plots for the analysis of teacher self-efficacy for the United States that was discussed earlier. As with the analysis of teacher job satisfaction, the analysis of teacher self-efficacy also shows some issues of convergence, however, the *Rhat* and n_eff (Table 2) indicate that convergence has been achieved. Again, on the basis of the posterior predictive *p*-value of 0.23, the model shows quite poor prediction of the United States' average teacher self-efficacy (Figure A A.8).

Figure 2 depicts the relationship between induction participation at a teacher's current school and teacher self-efficacy, controlling for teachers' gender and years of experience as a teacher. More detailed results can be found in Table 2.

Figure 2. Participation in any induction activities at current school and teachers' self-efficacy



Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Table 2. Participation in any induction activities predicting teachers' self-efficacy

Country/economy	Effective sample size	Rhat	Posterior mean (sd)	95% CI		$P(\lvert effect \rvert) \neq 0$	Original results[1]
				2.5%	97.5%		
United Arab Emirates	959	1	0.49 (0.09)	0.33	0.66	0.50	**0.48**
CABA (Argentina)[2]	900	1	0.47 (0.17)	0.13	0.77	0.50	**0.40**
Spain	794	1	0.40 (0.09)	0.24	0.56	0.50	**0.39**
Romania	1 007	1	0.34 (0.13)	0.10	0.59	0.50	**0.37**
Chile	961	1	0.41 (0.18)	0.05	0.75	0.49	**0.39**
Latvia	1 047	1	0.30 (0.12)	0.06	0.55	0.49	**0.29**
Czech Republic	855	1	0.25 (0.09)	0.09	0.43	0.49	**0.27**
Croatia	1 063	1	0.24 (0.10)	0.04	0.45	0.49	**0.27**

Country/economy	Effective sample size	Rhat	Posterior mean (sd)	95% CI		$P(\lvert effect\rvert) \neq 0$	Original results[1]
				2.5%	97.5%		
Colombia	1 184	1	0.23 (0.10)	0.02	0.43	0.49	0.10
Hungary	1 006	1	0.23 (0.10)	0.04	0.43	0.49	**0.24**
Kazakhstan	941	1	0.20 (0.09)	0.02	0.36	0.49	**0.28**
England (UK)	845	1	0.36 (0.18)	0.02	0.71	0.48	**0.36**
Estonia	996	1	0.25 (0.12)	0.01	0.49	0.48	**0.28**
Belgium	801	1	0.19 (0.09)	0.02	0.38	0.48	**0.28**
Saudi Arabia	1 054	1	0.36 (0.19)	-0.01	0.73	0.47	**0.53**
Shanghai (China)	1 060	1	0.31 (0.16)	0.01	0.62	0.47	**0.31**
Türkiye	814	1	0.27 (0.14)	0.01	0.48	0.47	0.23
Netherlands	934	1	0.22 (0.13)	-0.03	0.47	0.46	**0.24**
Italy	883	1	0.20 (0.11)	-0.01	0.40	0.47	**0.21**
France	1 069	1	0.25 (0.16)	-0.07	0.56	0.44	**0.37**
Finland	1 084	1	0.23 (0.14)	-0.03	0.50	0.46	**0.25**
Mexico	1 000	1	0.20 (0.12)	-0.03	0.44	0.46	**0.22**
Portugal	996	1	0.12 (0.08)	-0.04	0.26	0.44	**0.12**
Brazil	921	1	0.20 (0.14)	-0.07	0.47	0.43	**0.18**
Sweden	1 044	1	0.19 (0.15)	-0.11	0.48	0.40	0.14
Slovenia	1 013	1	0.17 (0.14)	-0.09	0.43	0.40	0.17
Japan	932	1	-0.18 (0.14)	-0.46	0.09	0.40	-0.08
Singapore	1 090	1	-0.26 (0.20)	-0.68	0.14	0.40	-0.19
Australia	1 057	1	0.15 (0.13)	-0.10	0.41	0.37	0.13
Viet Nam	997	1	0.14 (0.12)	-0.09	0.37	0.37	**0.24**
South Africa	1 071	1	0.19 (0.18)	-0.15	0.56	0.35	0.01
Israel	1 094	1	0.17 (0.17)	-0.19	0.53	0.33	0.24
Slovak Republic	973	1	0.11 (0.12)	-0.13	0.36	0.32	0.14
Denmark	838	1	0.10 (0.12)	-0.13	0.32	0.30	0.08
Lithuania	917	1	0.10 (0.12)	-0.13	0.34	0.29	0.13
Korea	914	1	0.12 (0.17)	-0.22	0.44	0.27	0.10
Austria	1 045	1	-0.08 (0.11)	-0.29	0.14	0.26	-0.07
Georgia	863	1	0.16 (0.23)	-0.29	0.57	0.25	0.03
Norway	954	1	-0.04 (0.08)	-0.18	0.12	0.20	-0.04
Alberta (Canada)	891	1	0.09 (0.21)	-0.29	0.51	0.18	0.03
Malta	946	1	0.08 (0.20)	-0.32	0.48	0.16	0.21
New Zealand	1 057	1	-0.06 (0.17)	-0.38	0.28	0.14	0.03
United States	1 024	1	-0.04 (0.15)	-0.32	0.24	0.10	-0.22
Bulgaria	930	1	-0.01 (0.11)	-0.23	0.22	0.03	-0.07

Notes:
1. Statistically significant results from OECD (2020[2]) are reported in bold.
2. CABA (Argentina): Ciudad Autónoma de Buenos Aires, Argentina.
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

An inspection of Table 2 also reveals that, across the countries/economies, the effective sample size is nearly 1 000, indicating low autocorrelation. In addition, the *Rhat* values are 1.0, indicating convergence of the algorithm. Substantive interpretations for Table 2 follow the same logic as those discussed above for Table 1. Take, for example, Sweden. Here we find that zero is in the 95% credible interval and it is not statistically significant based on the frequentist *p*-value. However, the estimated probability that the effect being different from zero is 0.40, which might be considered relatively large. Again, this type of nuanced interpretation of the results is only possible via Bayesian inference. Bubble plots for the remaining analyses for teacher self-efficacy can be found in Annex C.

8. A proposed Bayesian workflow for ILSA analyses

Our analyses of teacher job satisfaction and teacher self-efficacy in Sections 6 and 7, respectively, suggest a possible workflow for a Bayesian analysis of large-scale educational data utilising non-informative or weakly informative priors. Our proposed workflow follows one proposed in Chapter 12 of Kaplan (forthcoming[5]) but, of course, other workflows are possible, depending on the extent of detail desired in reporting research results (Gelman et al., 2020[28]; OECD, 2020[2]). Moreover, there are certainly similarities between some the steps of this workflow and the steps that could be followed in a frequentist analysis of the same data. The steps of our workflow are as follows.

1. Specify the outcome and set of predictors of interest, taking special care to note the assumptions regarding the distribution of the outcome – e.g. is the outcome assumed to be normally distributed, or does the outcome perhaps follow some type of non-normal distribution, such as the logistic or Poisson distribution. Specifying simple Bayesian models for the moments of the distribution (e.g. mean and variance) and examining the sensitivity of different prior choices can be quite useful and provide a sense of the probability model that generated the outcome. For this report, the outcome variables are scales and were treated as normally distributed.

2. Specify the functional form of the relationship between the outcome and the predictors. For the analysis of ILSA data generally, this will most likely be a type of linear or generalised linear model, but more complex models are, of course, possible. Because this report is styled to represent the analyses that were conveyed in the original TALIS reports, we utilised linear models, treating each predictor separately. As discussed above, we fully recognise the biases that might occur in treating the predictors separately, but it is beyond the scope of this report to develop a full predictive model of the outcomes of interest. As an aside, it is important to note that there may be more than one model that could have plausibly generated the data. Keeping the problem of model uncertainty in the back of one's mind is quite important, depending on the goals of the analysis. We discuss the issue of model uncertainty in the conclusions section of the report.

3. Take note of the complexities of the data structure – e.g. are the data generated from a clustered sampling design? Are there sampling weights? Accounting for the complexities of the data structure can be handled by careful specification of a Bayesian hierarchical model. The use of sampling weights can be easily incorporated in *Stan*-based programs such as *rstanarm* (Goodrich et al., 2020[24]) and *brms* (Bürkner, 2017[29]), and we have utilised TALIS sampling weights for this report.

4. Decide on the prior distributions for all parameters in the model. These priors will be either non-informative, weakly informative, informative, or a mix of all three. In the case of policy-oriented reports, such as the TALIS reports, it may be desirable to employ non-informative or weakly informative priors. In the former case, non-informative priors do not have the potential of reflecting the researcher's personal opinions and instead let the data speak. The latter case of weakly informative priors can be used to help stabilise computations, but do not contain very much additional information. Because the goal of the present report is to mimic the reporting of a policy-relevant report on TALIS, we utilised non-informative or weakly informative priors.

5. After running the analysis, it is essential that the convergence criteria of the algorithm be checked. The basics of Bayesian computation, along with convergence criteria, can be found in Kaplan (forthcoming[5]). Note that results cannot be communicated unless there is overwhelming evidence from a variety of diagnostics that the algorithm converged. There are instances, however, where there may be contradictory evidence of convergence. For example, trace plots may appear fine, but *Rhat* values may be somewhat problematic. All attempts should be made to improve these diagnostics before communicating the results. In most cases, if the effective sample size (n_eff) and *Rhat* values are reasonable, then one can proceed with communicating the results. This is because these diagnostics together capture autocorrelation, mixing, and trend in the iterations.

6. Given evidence of computational convergence, and with the results in hand, posterior predictive checking is a necessary step in the Bayesian workflow. Posterior predictive checks can be set up to gauge overall model fit but, depending on the goals of the analysis, specific posterior predictive checks can be provided regarding fit of specific aspects of the posterior predictive distribution. Two examples include assessing whether the model fits the variance of the distribution, or whether the model fits specific quantiles of the distribution such as extreme values.

7. Following posterior predictive checks, a full description of the posterior distributions of the model parameters would be provided, including the mean, standard deviation, and posterior intervals of interest. Additional posterior intervals of substantive interest should be provided, such as the probability that the effect is greater than (or less than, if negative) zero, or the probability that the effect lies between two values of substantive importance. For this report, we provided probabilities that the effects of interest are different from zero.

8. Sensitivity analyses should be conducted, examining the impact of the choice of priors on the substantive results. Sensitivity analyses can include simply comparing the findings to the case where all priors are non-informative, or to the case where very small changes to the mean and variance of the prior distributions are made. Note again, that with large sample sizes such as those encountered in this report, it is likely that results will be robust to reasonable alternative prior distributions.

9. Finally, though it was not discussed in this report, it may be important to examine model uncertainty. Addressing model uncertainty is particularly crucial if the goal of an analysis is to develop a model with optimal predictive performance, perhaps to be used for forecasting trends. One might also wish to investigate the extent of model uncertainty if the analyst is specifying a number of different models.

9. Conclusions

It is beyond the scope this report to list all of the advantages of Bayesian methods over frequentist methods. A broader list of advantages can be found in Kaplan (forthcoming[5]) and Wagenmakers et al. (2008[30]), however we list a set of important advantages which have immediate relevance to this report, and to the analysis and reporting of ILSAs generally.

9.1. Summarising the Bayesian advantage

1. Bayesian inference is the only paradigm of statistics that allows for the quantification of epistemic uncertainty – that is, uncertainty regarding our knowledge about unknown parameters. This form of uncertainty is not only present in our knowledge of the parameters of interest, but also in the very models that are used to estimate those parameters. Central to Bayesian theory and practice is that the posterior probability intervals around parameter estimates are more accurate in the sense that these intervals will accurately reflect epistemic uncertainty, particularly in small sample size cases, and they will be similar to frequentist confidence intervals (though with an entirely different interpretation) in large sample size cases. Bayesian models will also demonstrate better predictive performance than frequentist models by accounting for uncertainty in both the parameters of models and the choice of models themselves, they are better calibrated to reality (Dawid, 1982[31]; Kaplan, 2021[32]).
2. Bayesian inference provides posterior predictive checks, which allow one to examine the fit of the model with reference to its predictive performance. For the two examples in this report, evidence for good predictive fit was lacking, which suggests that one should proceed with caution in interpreting the results. In the context of our analyses, this result is not surprising insofar as each predictor was taken one at a time, and the regression model was, no doubt, highly misspecified. Nevertheless, posterior predictive checking is an integral part of any Bayesian workflow.
3. In large samples, Bayesian approaches and frequentist approaches will converge to very similar values, though their interpretations are different. As noted above, frequentist parameters are treated as fixed and only uncertainty due to sampling variability can be estimated through reference to the estimate's standard error. Bayesian estimates are interpreted probabilistically, and this, arguably, provides a much richer interpretation than the simple decision of whether a parameter estimate is statistically significant or not. For this report, we highlighted how Bayesian estimates provide interesting probabilistic interpretations as we proceeded through the results.
4. Related to the third point, perhaps the major advantage of Bayesian inference of relevance to the analysis and reporting of ILSA data is that the analyst can summarise the entire posterior distribution of the effect – a consequence of treating parameters as random. Thus, not only can one provide, say, a 95% posterior interval for the effect but, indeed, any interval of interest. In our analysis, we examined the probability that the effect is greater than zero. Additionally, we might wish to calculate the probability that the true effect lies between any two substantively important intervals. It is important to note that this kind of analysis is simply not possible in a frequentist setting.

9.2. What else can a Bayesian perspective offer?

In Section 8, we laid out a proposed Bayesian workflow that could be used for the analysis of ILSA data, generally. Step 9 of that workflow discussed the possibility of assessing model uncertainty, particularly if the goal is to develop optimally predictive models for policy forecasting (Kaplan, 2021[32]; Kaplan and Huang, 2021[33]). Although not the focus of this report, we believe that a distinct advantage of the Bayesian perspective lies in its power to yield optimally predictive models under uncertainty. Three general methods can be employed for this task: (1) sparse regression models, (2) Bayesian model averaging, and (3), Bayesian stacking. In all three cases, the presumption is that a model is being considered that includes many predictor variables.

9.2.1. Sparse regression modelling

In the case of sparse regression modelling, the goal is variable selection, and the approach is to employ priors that shrink small effects close to zero, while leaving large effects relatively unchanged. Examples include the *ridge prior*, the *lasso prior*, and the *horseshoe prior*, to name only three. A fuller discussion of sparsity in statistical modelling can be found in Hastie, Tibshirani and Friedman (2009[34])

9.2.2. Bayesian model averaging

The difficulty with sparse regression modelling is that, in the end, one is left with a single model that is often interpreted as the model the investigator had in mind all along (Hoeting et al., 1999[35]). This ignores the problem of model uncertainty that occurs when not knowing what the true data generating model might be. An approach that can be used to address this issue is Bayesian model averaging, a method that has been considered in the literature for decades and applied to many fields.

The main idea behind Bayesian model averaging (BMA) is that if there are, say, q variables in a regression model, then there are 2^q possible models that could have generated the outcome (not including interaction terms). Bayesian model averages searches across the 2^q models and measures the fit of each model given the data. These measures of fit are used as weights, and a weighted combination of the regression coefficients for each model appearing in the set are used to create a BMA-weighted average of regression coefficients. Theory and practice have found that, under certain important assumptions, a model using BMA-weighted averaged coefficients has better long-run predictive performance than any single model chosen by the researcher. For a review with applications to ILSAs, see Kaplan (2021[32]). An application of BMA to the United States' National Assessment of Educational Progress (NAEP) trend data was given in Kaplan and Huang (2021[33]).

9.2.3. Bayesian stacking

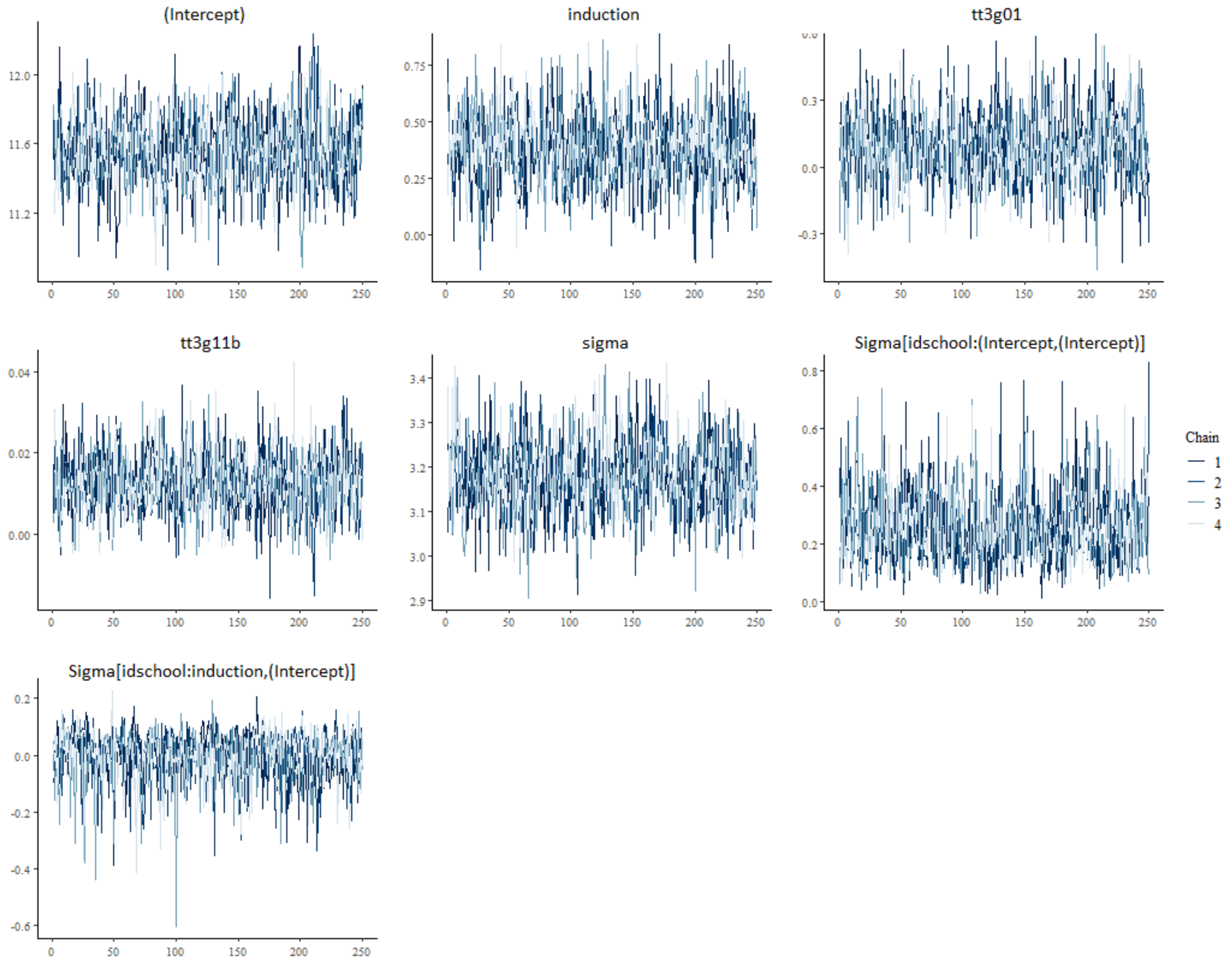
An important but technical limitation of BMA concerns the fact that its use requires the researcher to believe that the true data generating model is one of the 2^q models being explored by the method. That assumption is typically not likely to hold. To get around that problem, the method of Bayesian stacking holds promise. In Bayesian stacking, the researcher would specify a number of distinctly different models for the same outcome. These models could be based on policy or theoretical concerns, but in any case, the predictions from each model are obtained, and a mixture of the predictions are formed, where the mixture weights are based on the predictive quality of each model separately. Again, research suggests that Bayesian stacking performs at least as well, if not better than BMA in terms of predictive performance. A review of BMA and stacking with applications to the Programme for International Student Assessment (PISA) was given in Kaplan

(2021[32]). An application of BMA to the United States' National Assessment of Educational Progress (NAEP) trend data was given in Kaplan and Huang (2021[33]). A review and extension of Bayesian stacking to multilevel models with applications to PISA can be found in Huang and Kaplan (2023[36]).

To conclude, this report suggests an alternative approach to the analysis and reporting of TALIS data with relevance to other ILSAs. We attempted to stay close to the reporting style in OECD (2020[2]) while at the same time demonstrating key differences between the conventional significance testing approach in OECD (2020[2]) and the Bayesian alternative. Adopting the Bayesian alternative to analysis and reporting of TALIS, and ILSAs more generally, is not without some cost; perhaps, most importantly, considerable thought would need to be given regarding what constitutes substantively important effects. We recognise that this task is very difficult but maintain that it is still more beneficial to policy than simply providing an "up/down" significance test. Finally, we strongly recommend that additional consideration be given to predictive modelling described above.

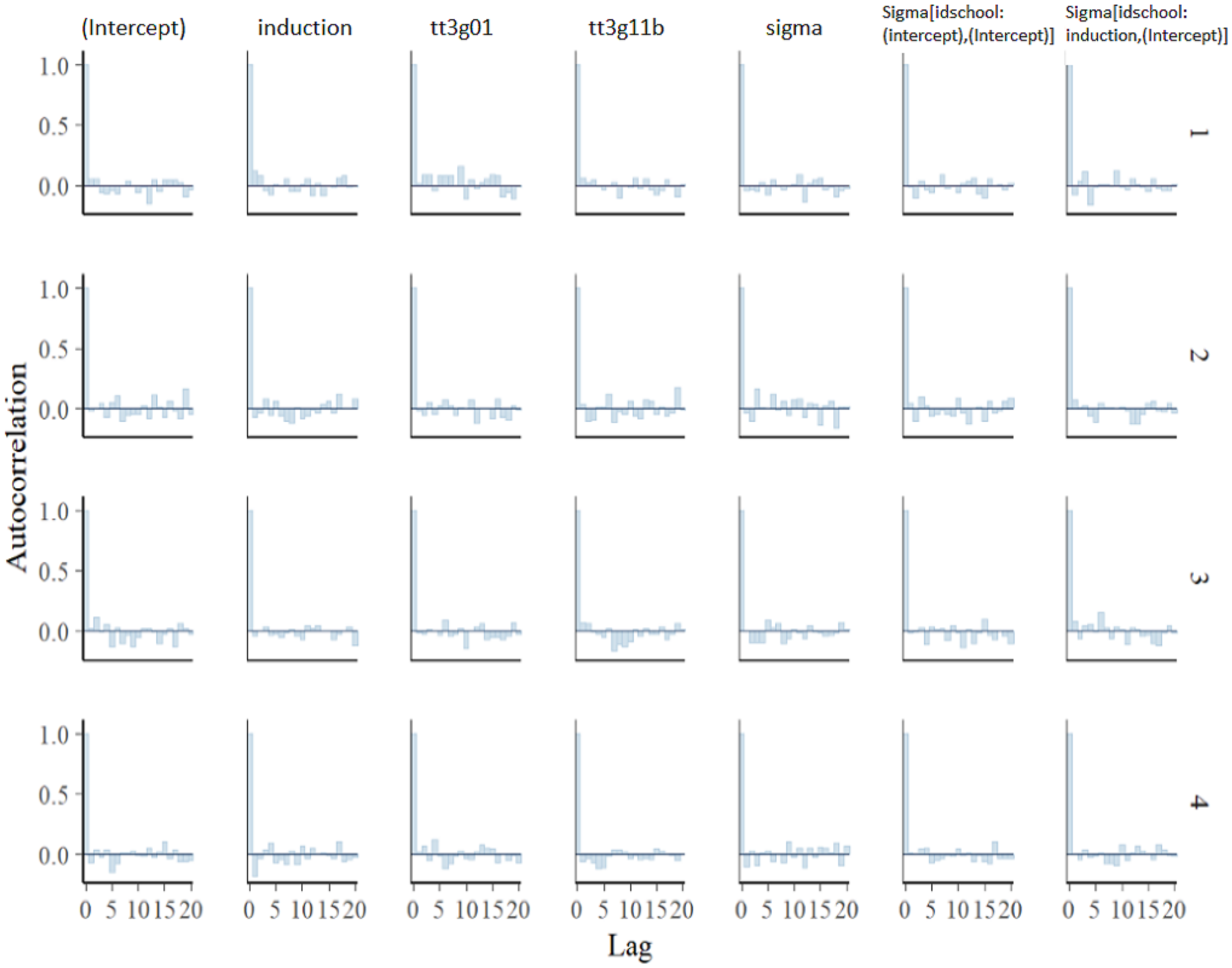
Annex A. Convergence plots

Figure A A.1. Analysis of teachers' job satisfaction – trace plots



Note: Trace plots for the model predicting teacher job satisfaction by participation in any induction activities at current school, United States sample. These plots should exhibit a clear rectangular horizontal band over the x-axis. These plots show some problems with the mixing of the chains.
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

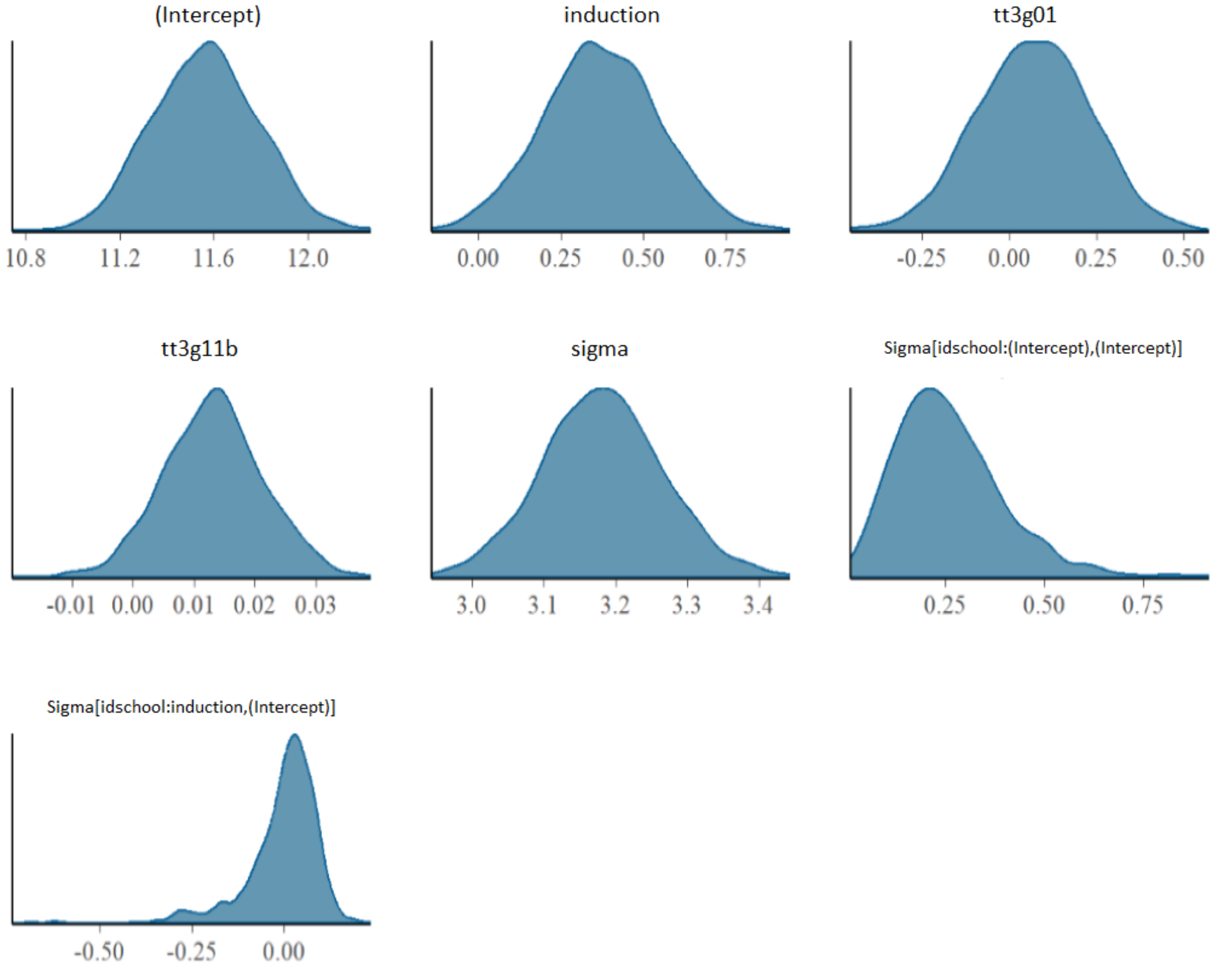
Figure A A.2. Analysis of teachers’ job satisfaction – autocorrelation plots



Note: Autocorrelation plots for the model predicting teacher job satisfaction by participation in any induction activities at current school, United States sample. These plots should show a very high autocorrelation at the first lag and very small autocorrelations thereafter. These plots show very low autocorrelation signifying independent draws from the posterior distributions.

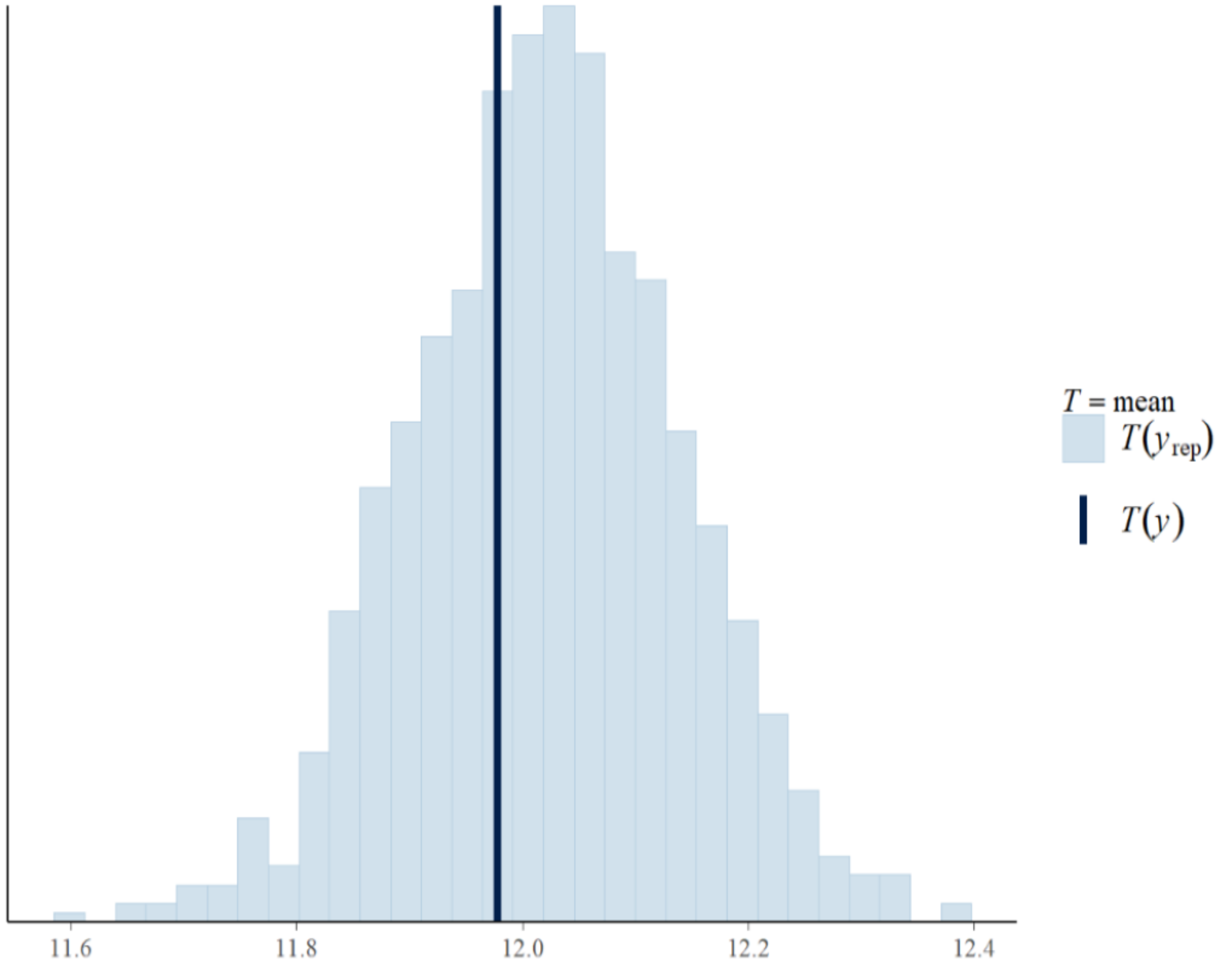
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A A.3. Analysis of teachers’ job satisfaction – density plots



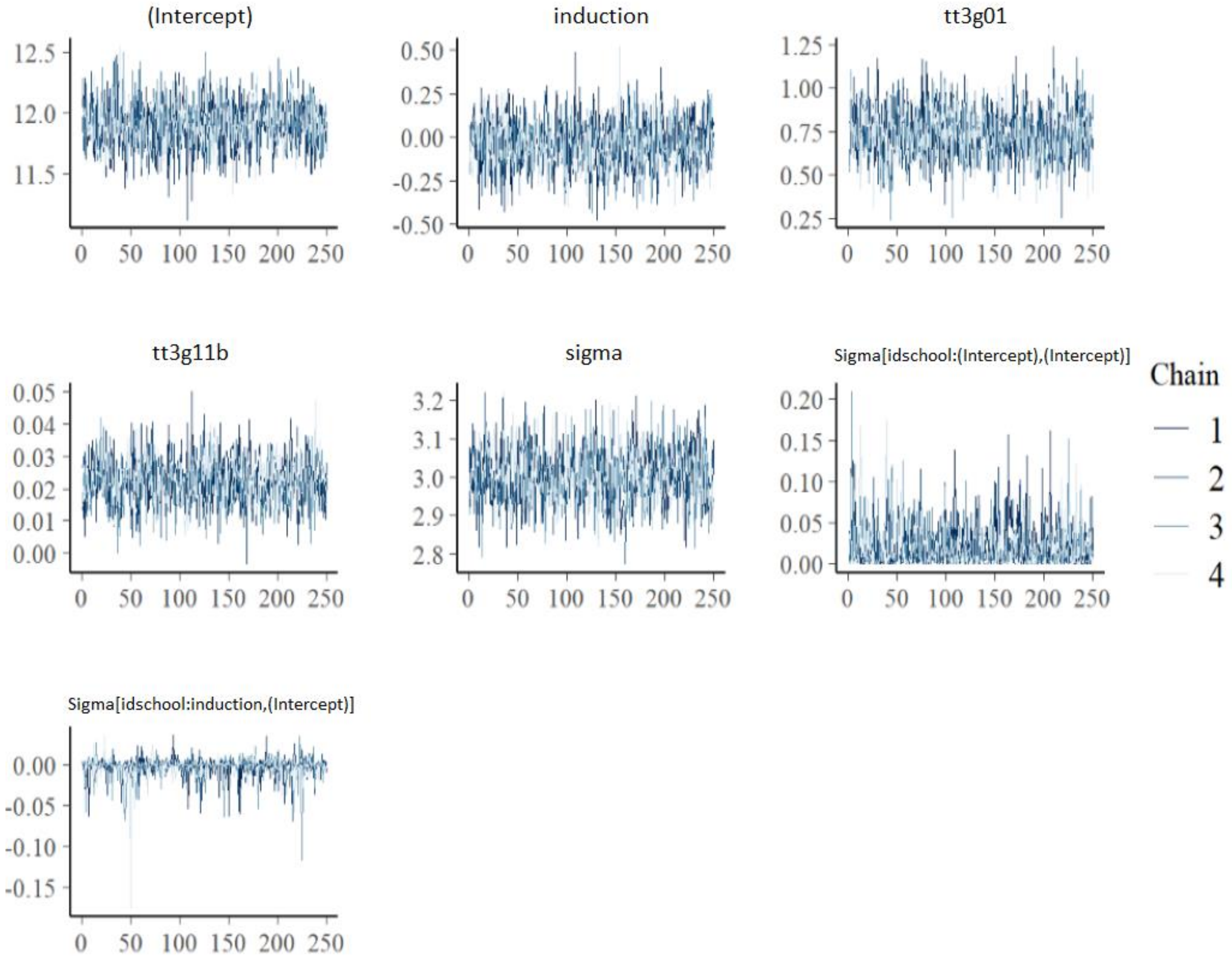
Note: Posterior probability distribution (density) plots for the model predicting teacher job satisfaction by participation in any induction activities at current school, United States sample. These plots should exhibit more or less a bell-shaped curve. We note some small problems with the variance components.
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A A.4. Analysis of teachers' job satisfaction – posterior predictive checks plots



Note: Posterior predictive check plots for the model predicting teacher job satisfaction by participation in any induction activities at current school (p = .65), United States sample. This plot should exhibit a bell-shaped curve with the test-statistic for the data (denoted by the solid black line) positioned at the centre of the distribution (0.50), indicating excellent fit.
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

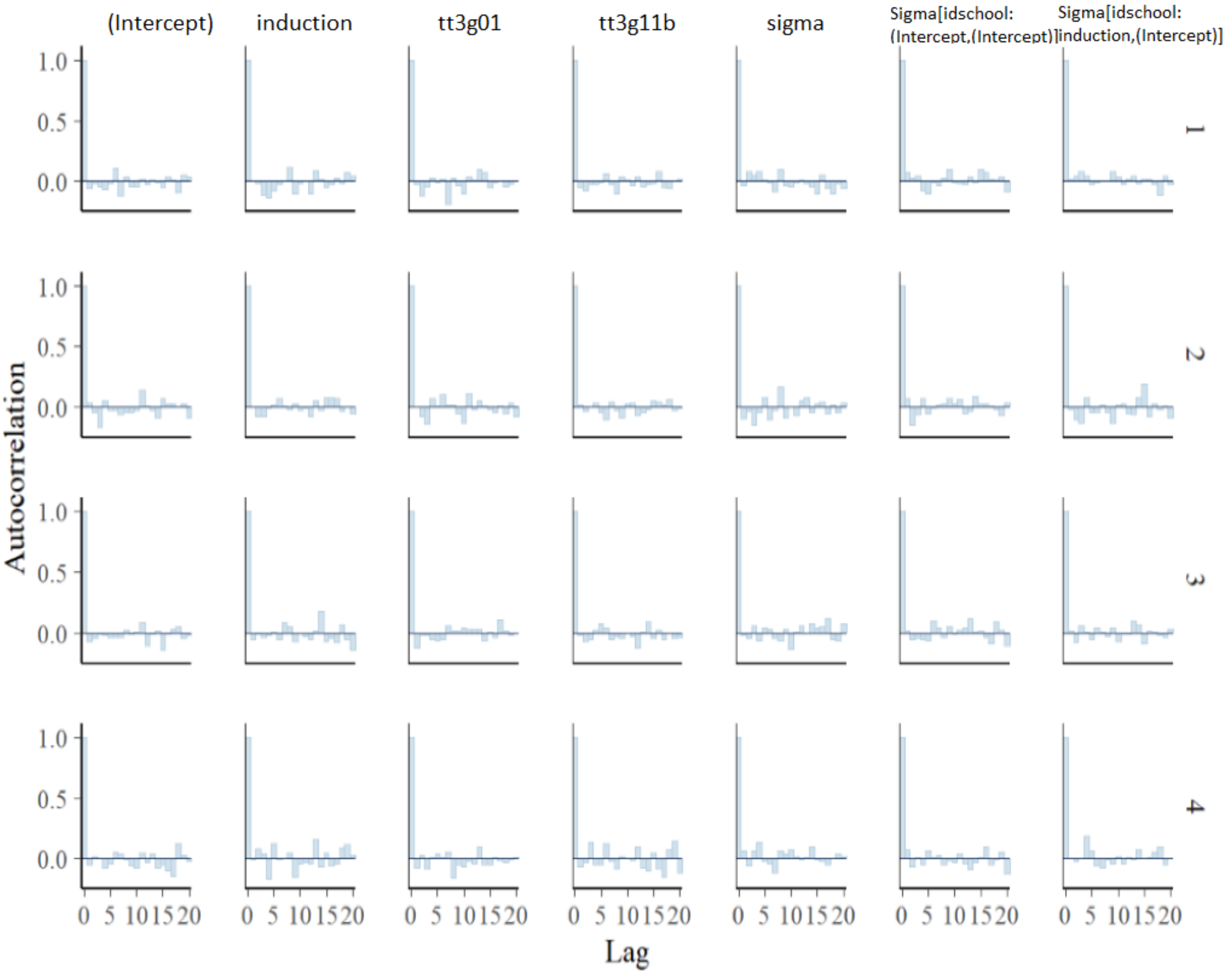
Figure A A.5. Analysis of teachers' self-efficacy – trace plots



Note: Trace plots for the model predicting teacher self-efficacy by participation in any induction activities at current school, United States sample. These plots should exhibit a clear rectangular horizontal band over the x-axis. These plots show some problems with the mixing of the chains.

Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

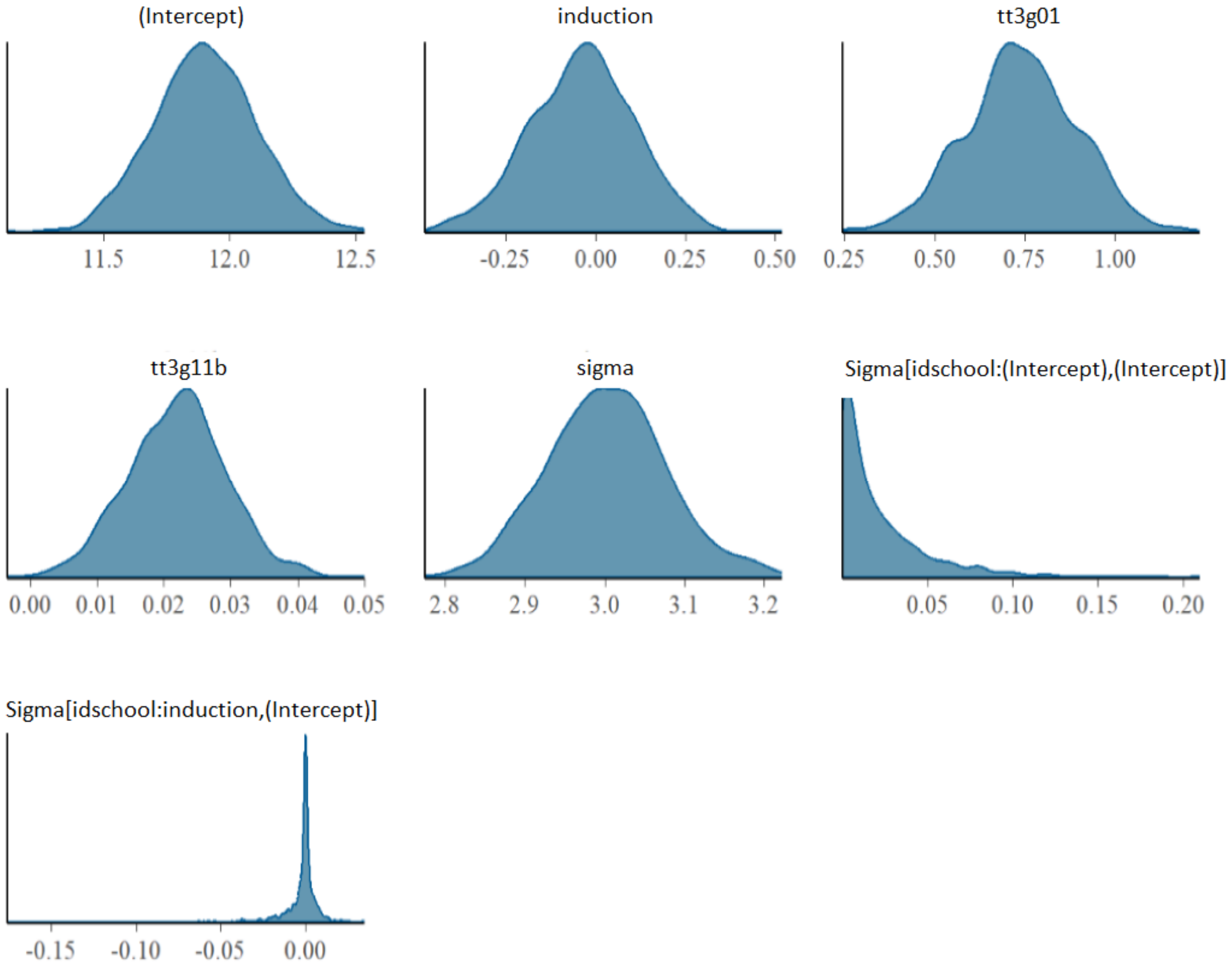
Figure A A.6. Analysis of teachers' self-efficacy – autocorrelation plots



Note: Autocorrelation plots for the model predicting teacher self-efficacy by participation in any induction activities at current school, United States sample. These plots should show a very high autocorrelation at the first lag and very small autocorrelations thereafter. These plots show very low autocorrelation signifying independent draws from the posterior distributions.

Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

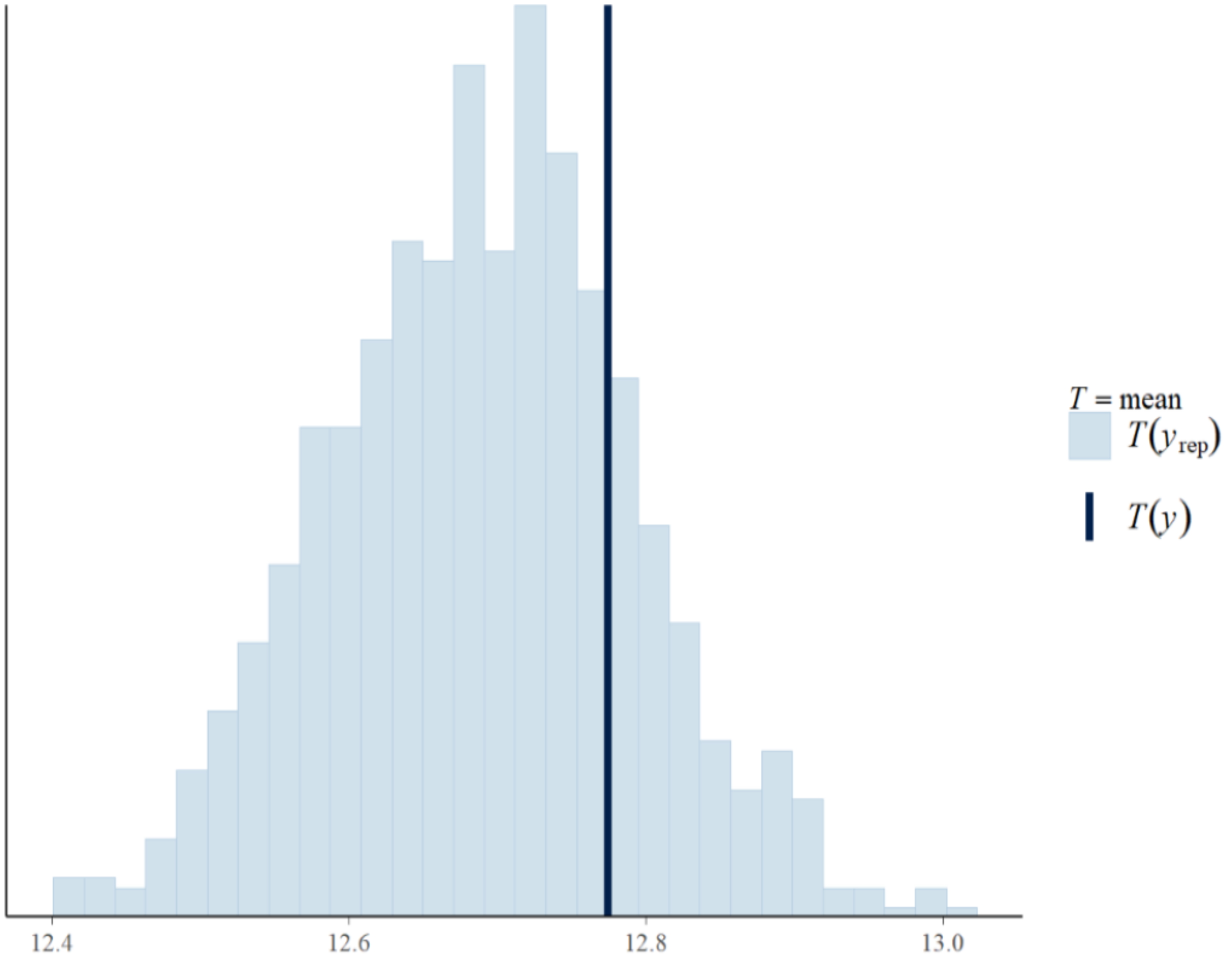
Figure A A.7. Analysis of teachers' self-efficacy – density plots



Note: Posterior probability distribution (density) plots for the model predicting teacher self-efficacy by participation in any induction activities at current school, United States sample. These plots should exhibit more or less a bell-shaped curve. We note some small problems with the variance components.

Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

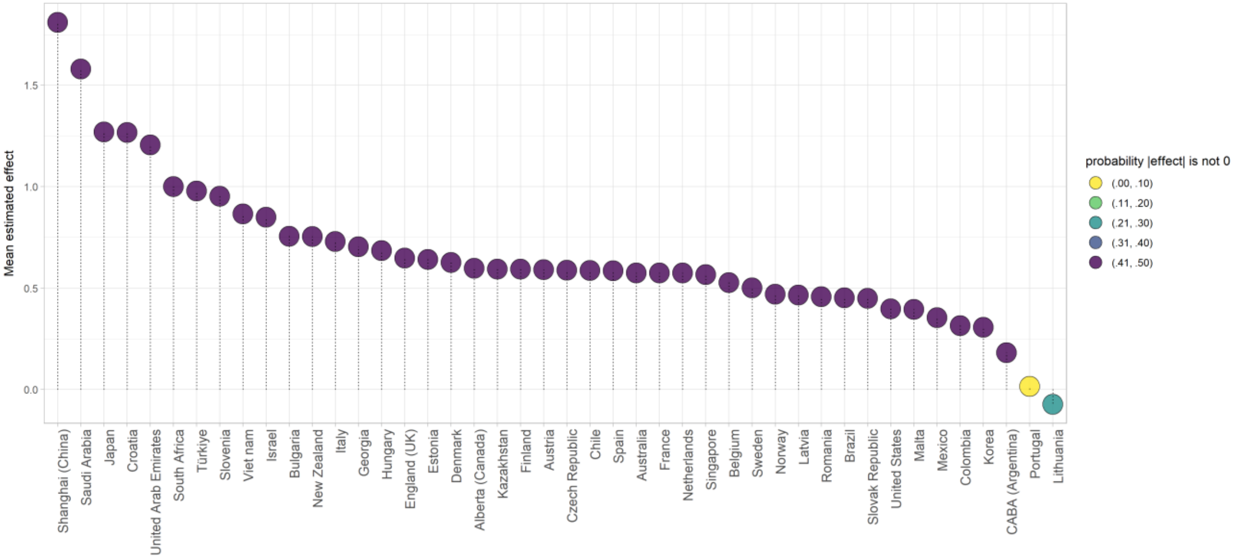
Figure A A.8. Analysis of teachers' self-efficacy – posterior predictive checks plots



Note: Posterior predictive check plots for the model predicting teacher self-efficacy by participation in any induction activities at current school (p = .20), United States sample. This plot should exhibit a bell-shaped curve with the test-statistic for the data (denoted by the solid black line) positioned at the centre of the distribution (0.50), indicating excellent fit.
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

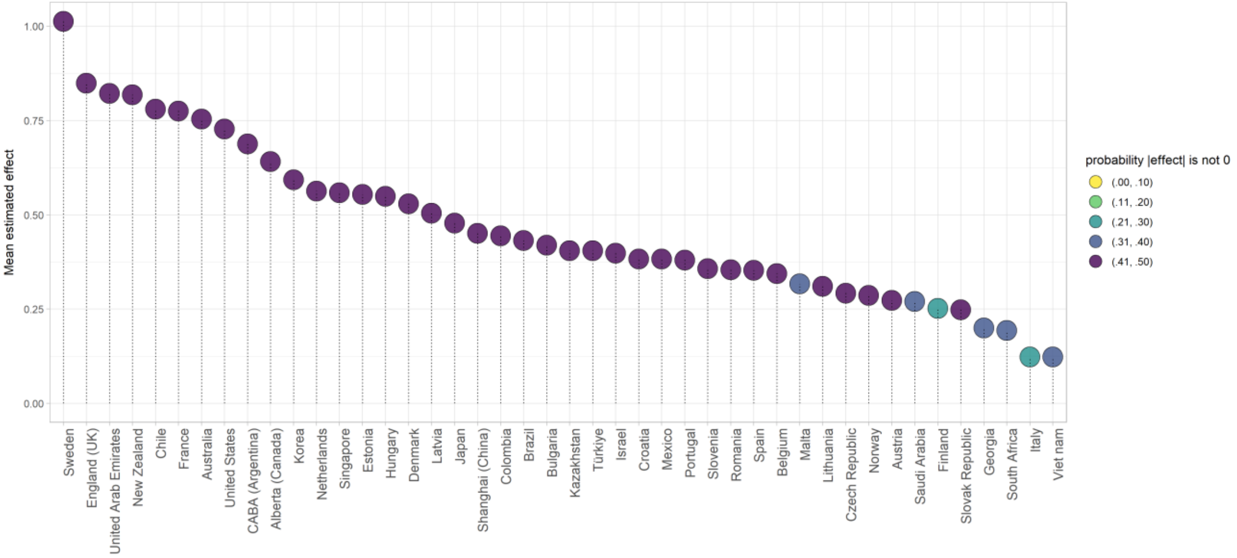
Annex B. Additional results for teachers' job satisfaction

Figure A B.1. Teaching as a first career choice



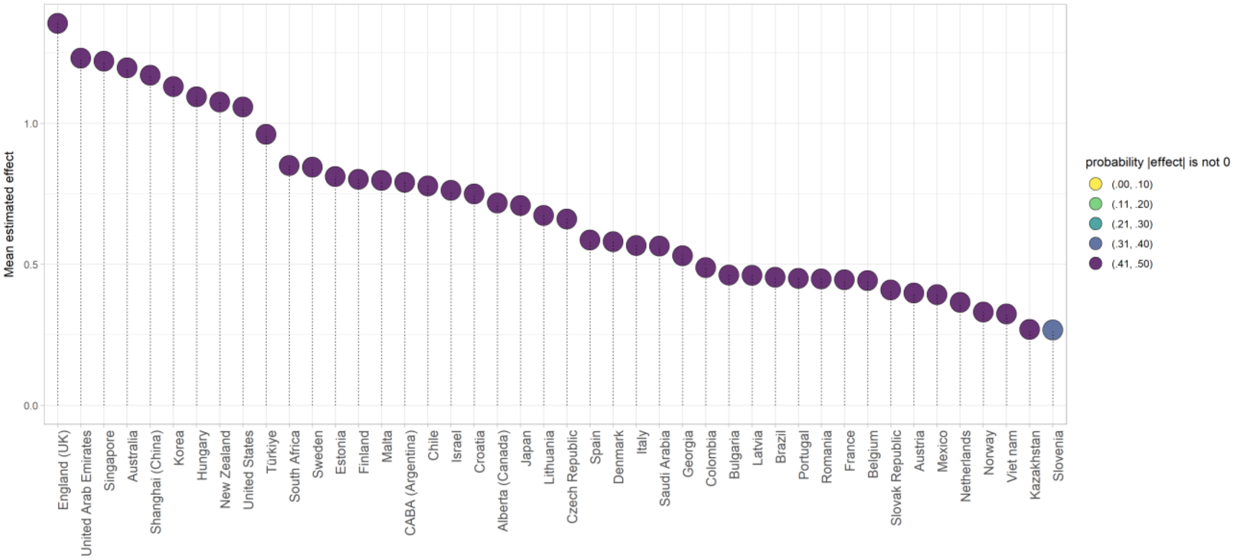
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A B.2. Induction at current school included team teaching with experienced teachers



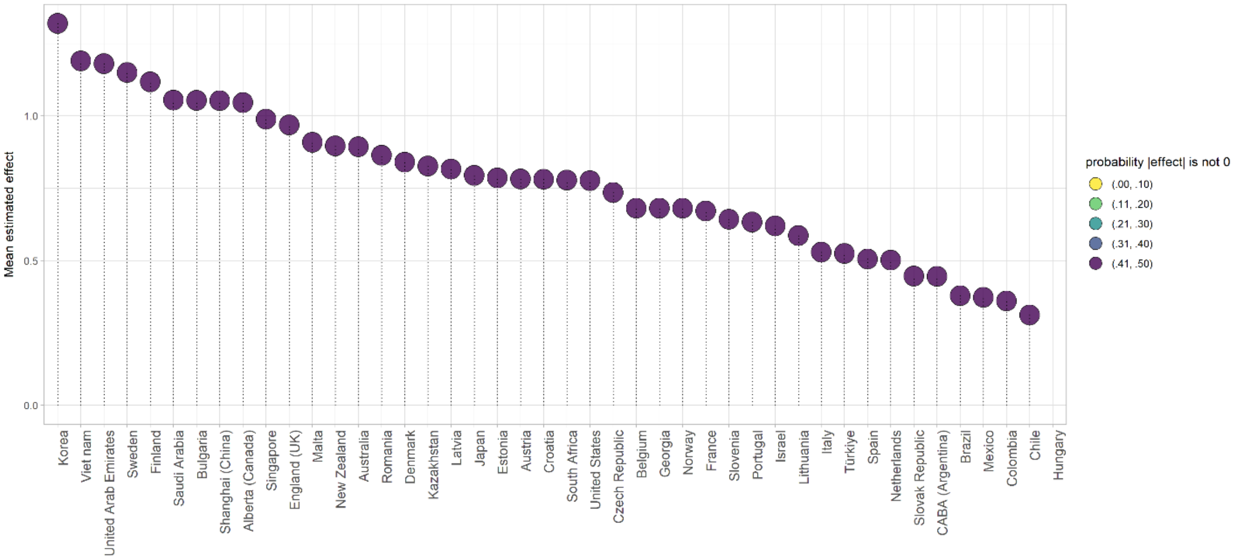
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A B.3. Professional development activities had a positive impact on teaching practices



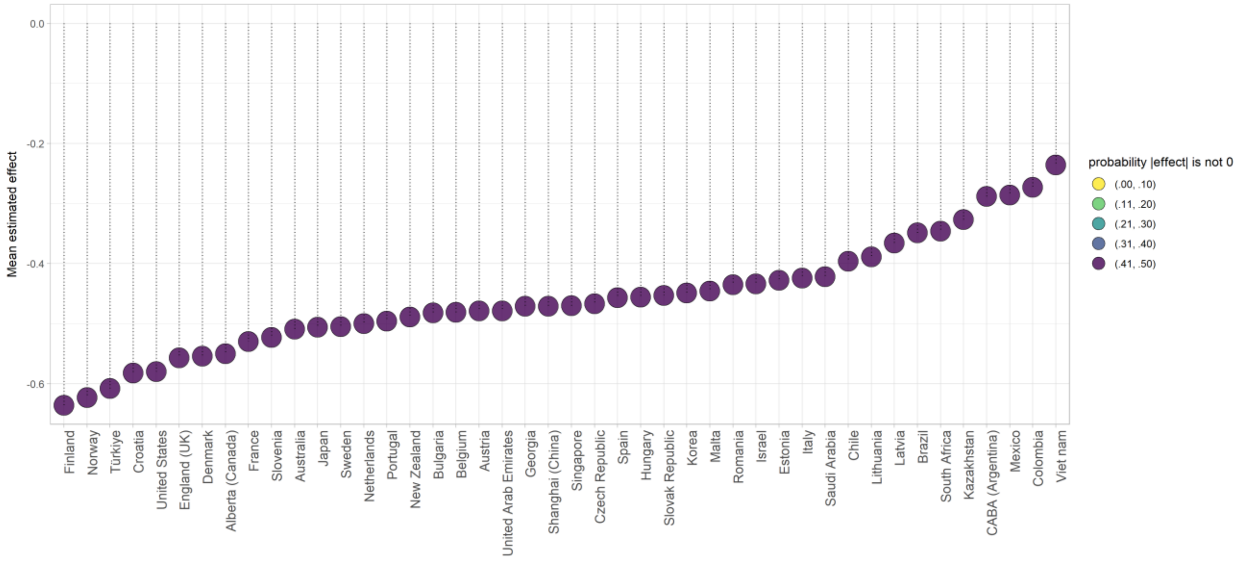
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A B.4. Teaching profession is valued in society



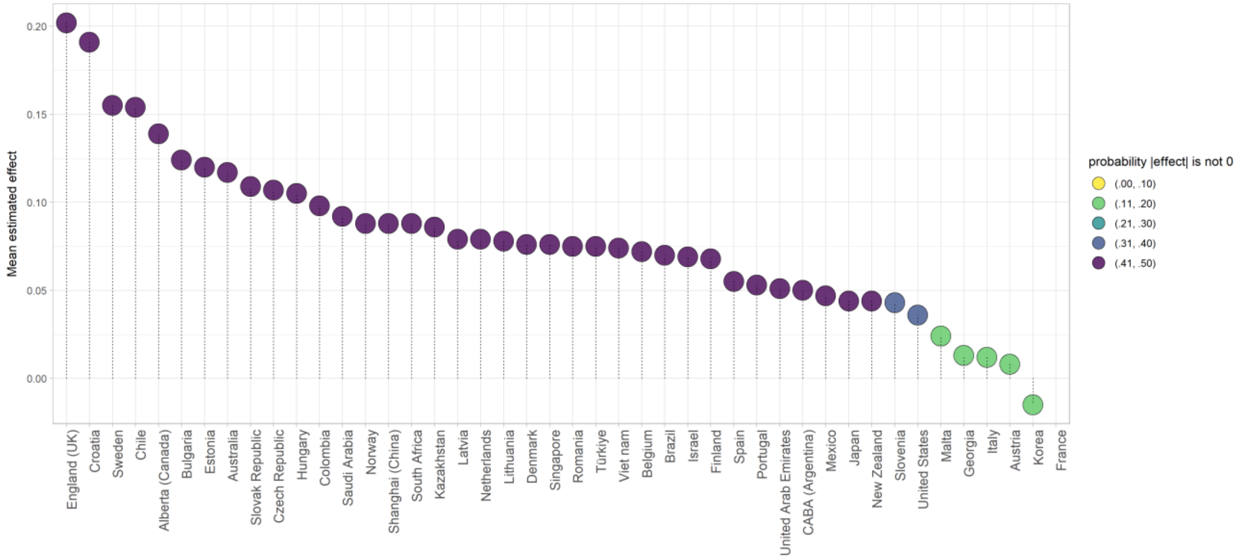
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A B.5. Index of workplace well-being and stress



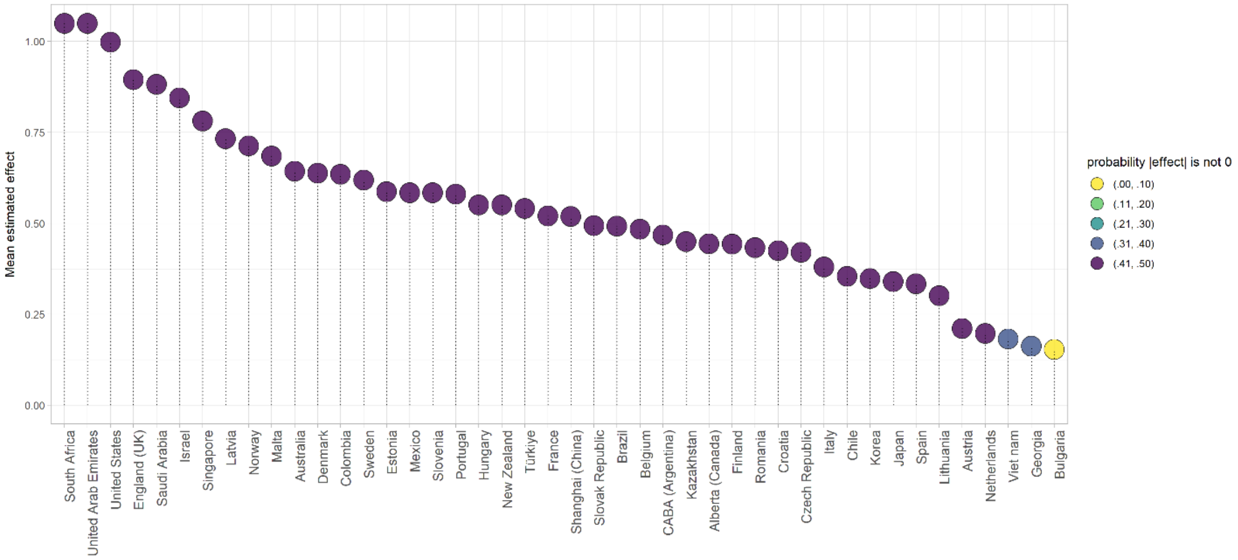
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A B.6. Index of professional collaboration



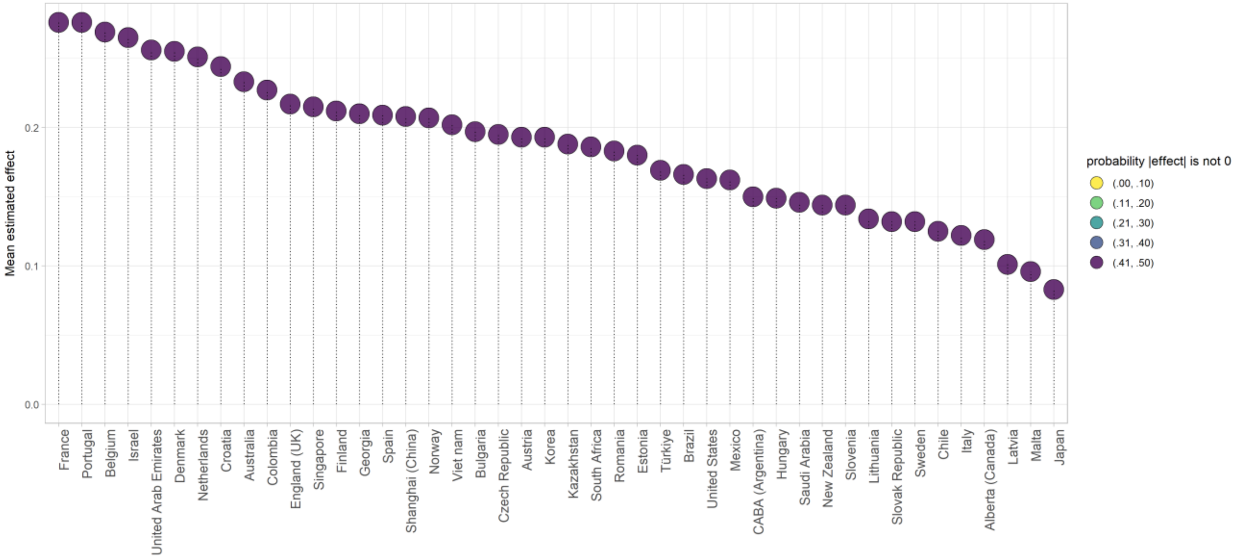
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A B.7. Receiving impactful feedback



Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

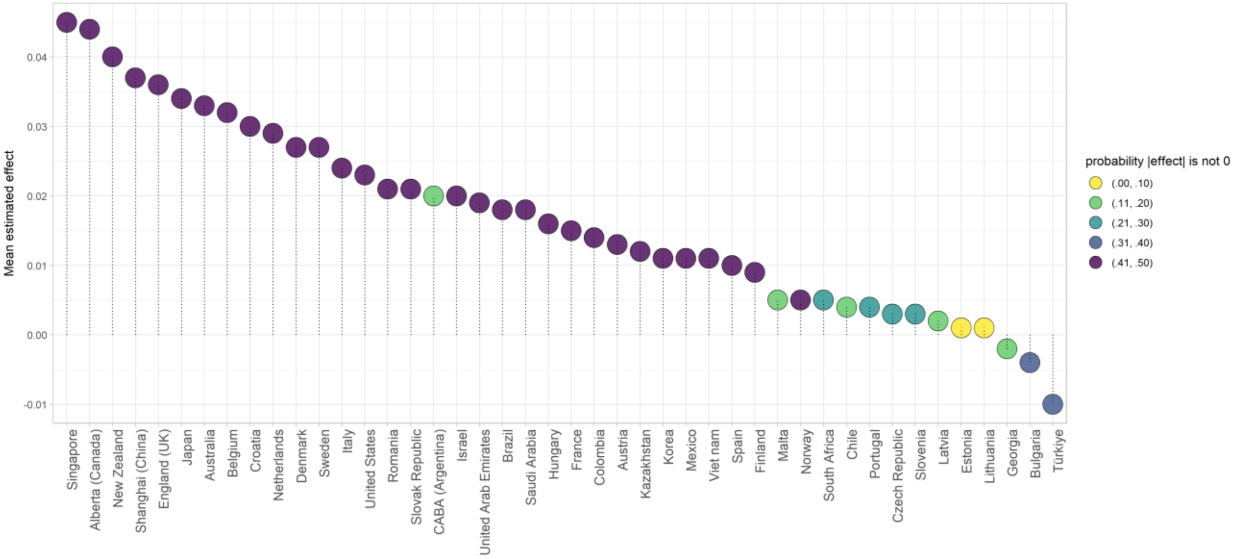
Figure A B.8. Index of autonomy in the target class



Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

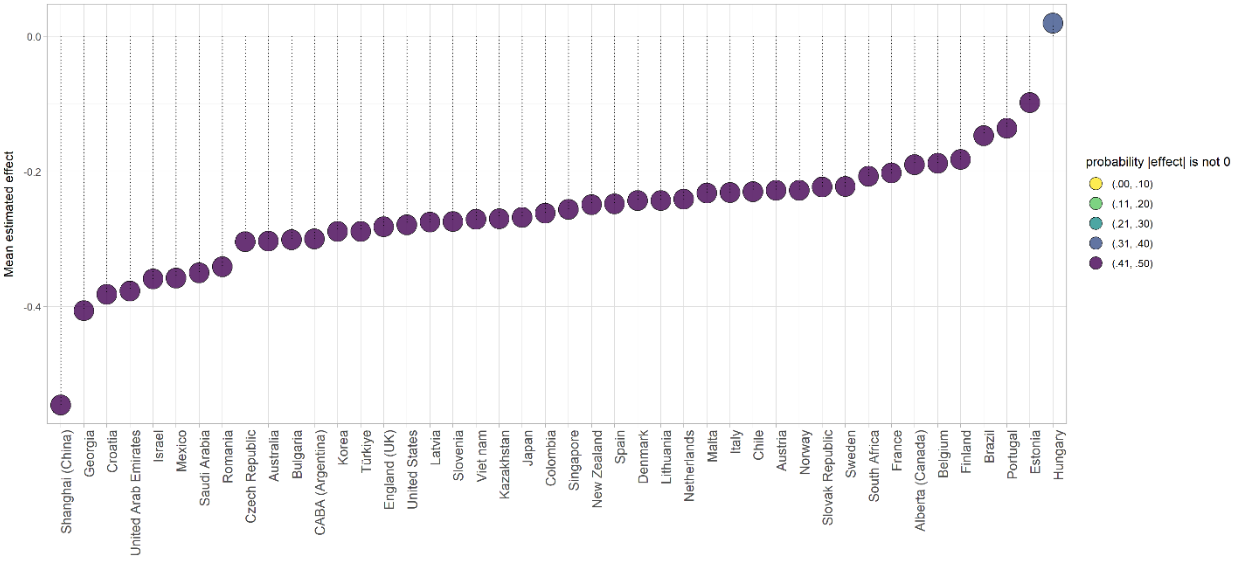
Annex C. Additional results for teachers’ self-efficacy

Figure A C.1. Years of experience as a teacher



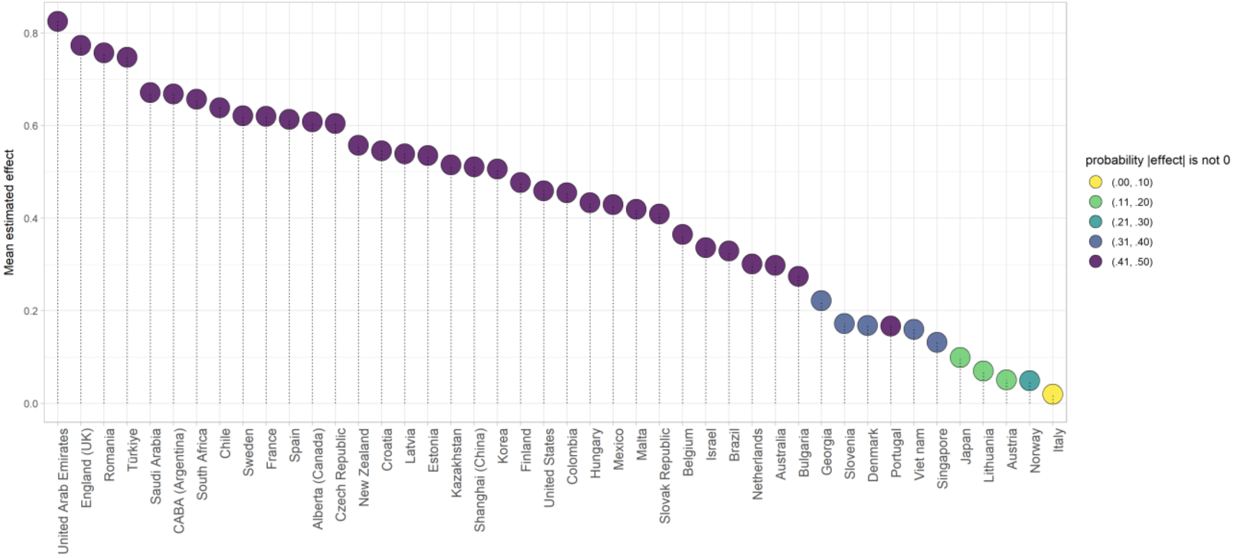
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A C.2. Index of classroom disciplinary climate



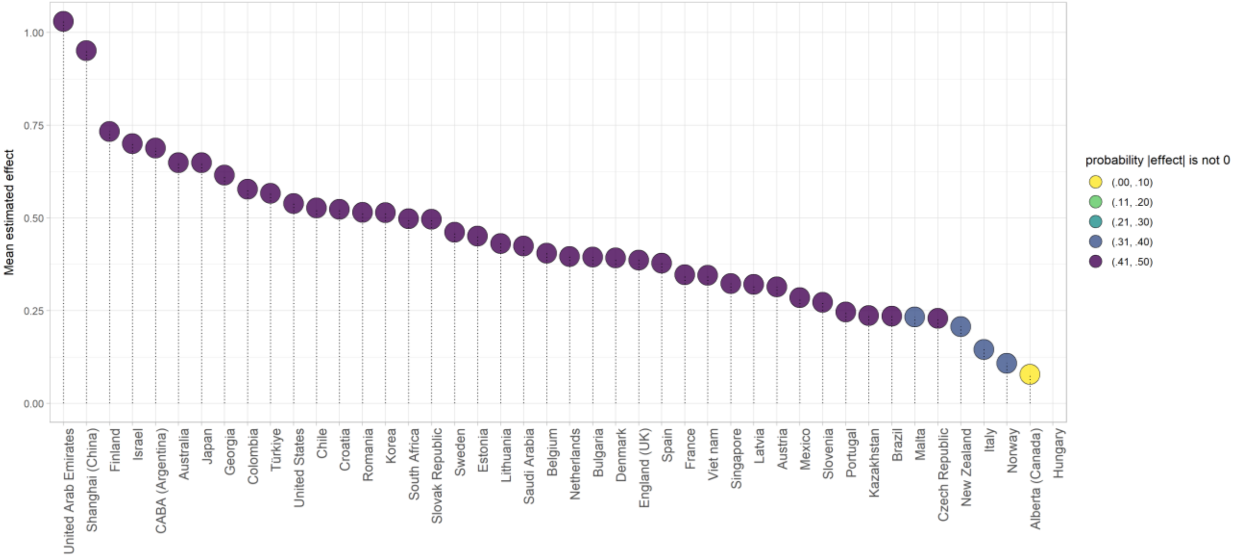
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A C.3. Induction at current school included team teaching with experienced teachers



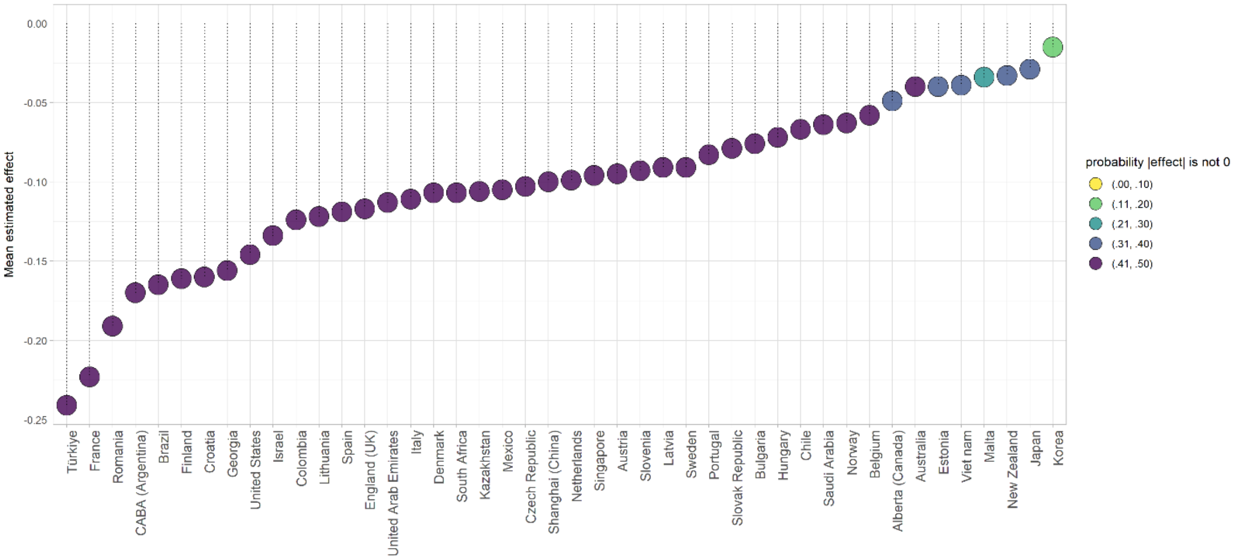
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A C.4. Professional development in the last 12 months had a positive impact on teaching practices



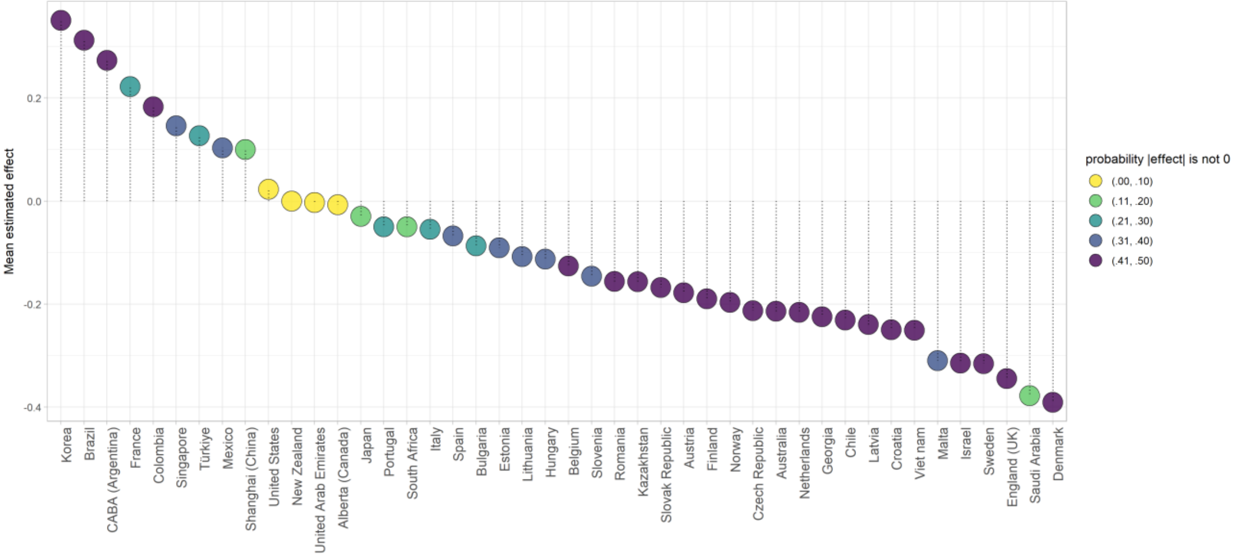
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A C.5. Index of workplace well-being and stress



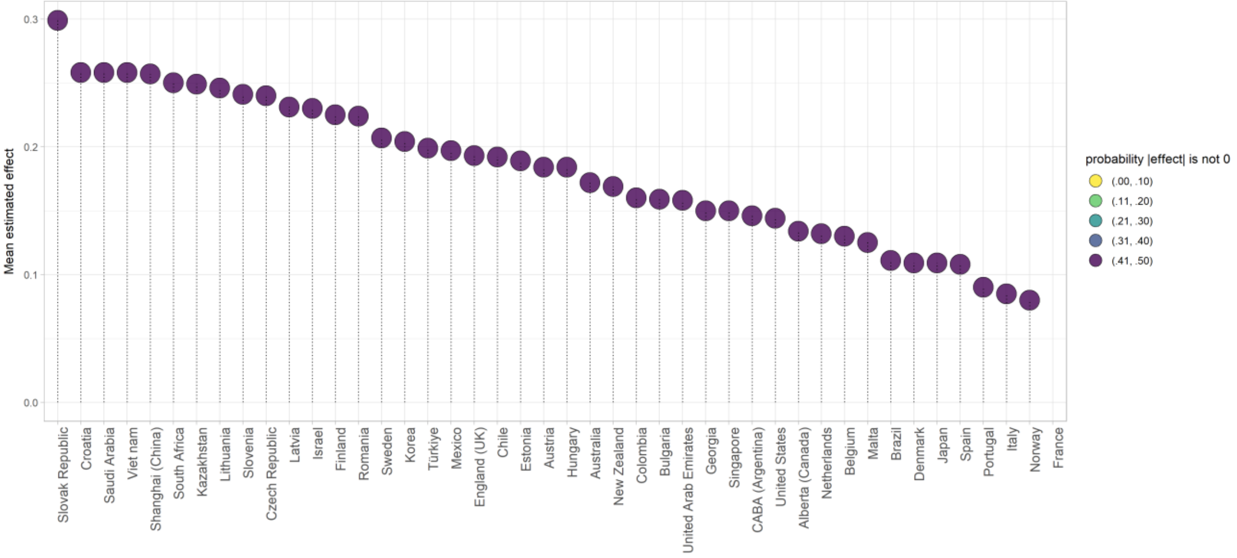
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A C.6. Fixed-term contract: less than one school year



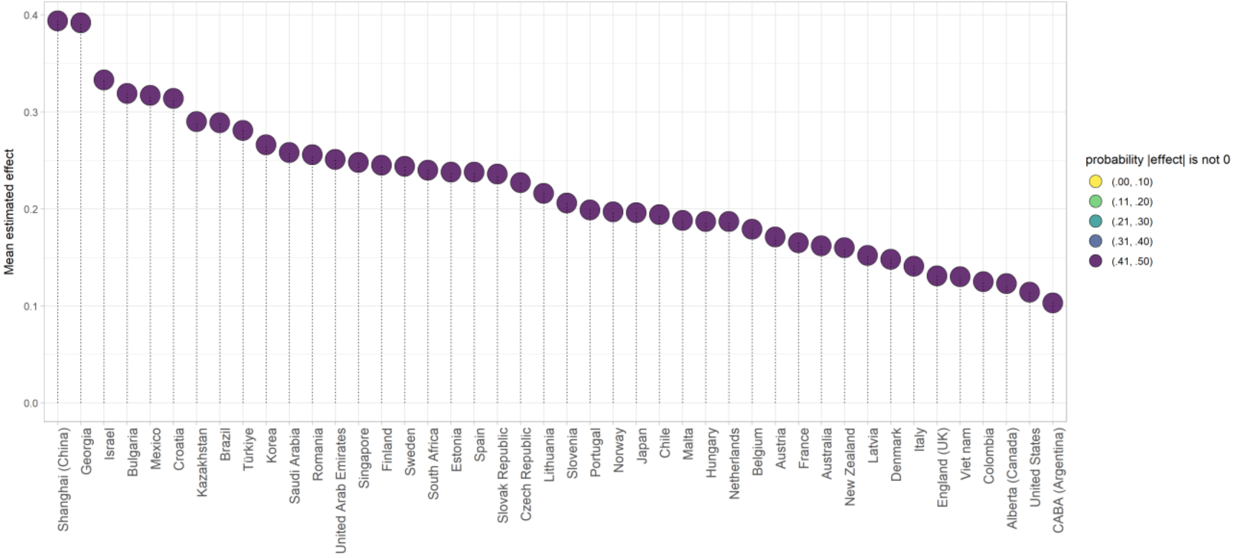
Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A C.7. Index of professional collaboration



Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

Figure A C.8. Index of autonomy in the target class



Source: OECD (2018[27]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022).

References

Bayes, T. (1763), “LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S., communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S”, *Philosophical Transactions of the Royal Society of London*, Vol. 53, pp. 370-418, https://doi.org/10.1098/rstl.1763.0053. [12]

Bürkner, P. (2017), “brms: An R package for Bayesian multilevel models using Stan”, *Journal of Statistical Software*, Vol. 80, pp. 1-28, https://doi.org/doi:10.18637/jss.v080.i01. [29]

Burstein, L. (1980), “Chapter 4: The analysis of multilevel data in educational research and evaluation”, *Review of Research in Education*, Vol. 8/1, pp. 158-233, https://doi.org/10.3102/0091732X008001158. [19]

Dawid, A. (1982), “The well-calibrated Bayesian”, *Journal of the American Statistical Association*, Vol. 77, pp. 605-610, https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477856. [31]

De Finetti, B. (1974), *Theory of Probability: A Critical Introductory Treatment. Volume 1*, John Wiley & Sons, New York, NY. [8]

De Finetti, B. (1974), *Theory of Probability: A Critical Introductory Treatment. Volume 2*, John Wiley & Sons, New York, NY. [9]

Gelman, A. (2006), “Prior distributions for variance parameters in hierarchical models”, *Bayesian Analysis*, Vol. 1, pp. 515-534, https://doi.org/10.1214/06-BA117A. [14]

Gelman, A. et al. (2014), *Bayesian Data Analysis, Third Edition*, Chapman and Hall, London. [4]

Gelman, A. and C. Shalizi (2012), “Philosophy and the Practice of Bayesian Statistics”, *British Journal of Mathematical and Statistical Psychology*, Vol. 66/1, pp. 8-38, https://doi.org/10.1111/j.2044-8317.2011.02037.x. [18]

Gelman, A. et al. (2020), “Bayesian workflow”, *arXiv*, https://doi.org/10.48550/ARXIV.2011.01808. [28]

Geman, S. and D. Geman (1984), “Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6/6, pp. 721-741, https://doi.org/10.1109/TPAMI.1984.4767596. [15]

Goldstein, H. (2011), *Multilevel Statistical Models, 4th Edition*, Wiley Series in Probability and Statistics, Wiley, New York, NY. [20]

Goodrich, B. et al. (2020), *rstanarm: Bayesian applied regression*, https://mc-stan.org/rstanarm. [24]

Hastie, T., R. Tibshirani and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Series in Statistics, Springer, New York NY. [34]

Hoeting, J. et al. (1999), “Bayesian model averaging: A tutorial”, *Statistical Science*, Vol. 14/4, pp. 382-417, https://www.jstor.org/stable/2676803. [35]

Howson, C. and P. Urbach (2006), *Scientific Reasoning: The Bayesian Approach*, Open Court, Chicago, IL. [11]

Huang, M. and D. Kaplan (2023), “Bayesian stacking in multilevel models”, *PsyArXiv* 21 March, https://doi.org/10.31234/osf.io/e9m6x. [36]

Kaplan, D. (2021), “On the quantification of model uncertainty: A Bayesian perspective”, *Psychometrika*, Vol. 86/1, pp. 215-238, https://doi.org/10.1007/s11336-021-09754-5. [32]

Kaplan, D. (forthcoming), *Bayesian Statistics for the Social Sciences*, Guilford Press, New York. [5]

Kaplan, D. and M. Huang (2021), “Bayesian probabilistic forecasting with large-scale educational trend data: A case study using NAEP”, *Large-scale Assessments in Education*, Vol. 9/15, https://doi.org/10.1186/s40536-021-00108-2. [33]

Kolmogorov, A. (1956), *Foundations of the Theory of Probability*, Chelsea, New York, NY. [7]

Laplace, P. (1951), *A Philosophical Essay on Probabilities*, Translated from the 6th French Edition by F. W. Truscott and F. L. Emory, Dover, New York, NY. [13]

Lindley, D. (2007), *Understanding Uncertainty*, Wiley, New York, NY. [10]

Metropolis N., A. et al. (1953), “Equations of state calculations by fast computing machines”, *Journal of Chemical Physics*, Vol. 21/6, pp. 1087-1091, https://doi.org/10.1063/1.1699114. [16]

OECD (2020), *TALIS 2018 Results (Volume II): Teachers and School Leaders as Valued Professionals*, TALIS, OECD Publishing, Paris, https://doi.org/10.1787/19cf08df-en. [2]

OECD (2019), *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners*, TALIS, OECD Publishing, Paris, https://doi.org/10.1787/1d0bc92a-en. [1]

OECD (2018), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 28 July 2022). [27]

Raudenbush, S. and A. Bryk (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition*, Sage Publications, Thousands Oaks, CA. [21]

Robert, C. and G. Casella (2011), “A short history of Markov Chain Monte Carlo: Subjective recollections from incomplete data”, *Statistical Science*, Vol. 26/1, pp. 102-115, https://www.jstor.org/stable/23059158. [17]

Rubin, D. (1986), “Statistical matching using file concatenation with adjusted weights and multiple imputation”, *Journal of Business & Economic Statistics*, Vol. 4/1, pp. 87-94, https://doi.org/10.1080/07350015.1986.10509497. [22]

Savage, L. (1954), *The Foundations of Statistics*, John Wiley & Sons, New York, NY. [37]

Stan Development Team (2021), *Stan Reference Manual: Version 2.26*, https://mc-stan.org/docs/2_26/reference-manual-2_26.pdf. [25]

Stan Development Team (2021), *Stan User's Guide: Version 2.26*, https://mc-stan.org/docs/2_26/stan-users-guide-2_26.pdf. [26]

van Buuren, S. (2012), *Flexible Imputation of Missing Data*, Chapman & Hall, New York, NY, https://doi.org/10.1201/b11826. [23]

Wagenmakers, E. (2007), "A practical solution to the pervasive problems of p values", *Psychonomic Bulletin & Review*, Vol. 14, pp. 779-804, https://doi.org/10.3758/BF03194105. [6]

Wagenmakers, E. et al. (2008), "Bayesian versus frequentist inference", in Hoijtink, H., Klugkist I and P. Boelen (eds.), *Bayesian Evaluation of Informative Hypotheses: Statistics for Social and Behavioral Sciences*, Springer, New York, NY, https://doi.org/10.1007/978-0-387-09612-4_9. [30]

Wasserstein, R. and N. Lazar (2016), "The ASA Statement on p-Values: Context, process, and purpose", *The American Statistician*, Vol. 70/2, pp. 129-133, https://doi.org/10.1080/00031305.2016.1154108. [3]