Kaplan and Harra
Large-scale Assessments in Education (2024) 12:2
https://doi.org/10.1186/s40536-023-00189-1

Large-scale Assessments in Education

METHODOLOGY Open Access

A Bayesian workflow for the analysis and reporting of international large-scale assessments: a case study using the OECD teaching and learning international survey

David Kaplan[1*] and Kjorte Harra[1]

*Correspondence: david.kaplan@wisc.edu

[1] University of Wisconsin - Madison, Madison, USA

Abstract

This paper aims to showcase the value of implementing a Bayesian framework to analyze and report results from international large-scale assessments and provide guidance to users who want to analyse ILSA data using this approach. The motivation for this paper 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 assessments more specifically. The paper 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 assessment reports. The data utilized for this paper comes from the Teaching and Learning International Survey (TALIS). The paper provides steps in implementing a Bayesian analysis and proposes a workflow that can be applied not only to TALIS but to large-scale assessments in general. The paper closes with a discussion of other Bayesian approaches to international large-scale assessment data, in particularly for predictive modeling.

Introduction

This paper aims to showcase the value of implementing a Bayesian framework to analyse and report on data from international large-scale assessments (ILSAs) with the OECD Teaching and Learning International Survey (OECD, 2019, 2020) serving as an example, and to provide guidance to users who want to analyse ILSA data using this approach. The motivation for this paper 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 assessments 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 analysts assumptions, before seeing the data, as to the expected value and variance of a parameter. The analysts prior assumptions can be more or less “informative”, arising from a summary of past research, expert opinion, or both. The mechanics of Bayes’ 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 parameters, Bayesian analysis provides a 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 more nuanced analysis of the effects of interest, and is, arguably, more informative to policy makers than simply indicating whether an effect is statistically significant or not.

Purpose and organization of paper

The OECD published a two-volume report based on the results of TALIS 2018. The first volume was entitled *TALIS 2018 Results: Teachers and school leaders as life long learners* (OECD, 2019) and the second volume was entitled *TALIS 2018 Results: Teachers and school leaders as valued professionals* (OECD, 2020). 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 summarizes the results of various regression analyses aimed at identifying relevant predictors of teacher job-satisfaction and teacher self-efficacy separately (See Figures II.1.7 and II.1.8 in OECD (2020)). The analyses of these outcomes were carried out as follows. For each country, least-squares regression analysis was conducted with the TALIS composite scales of teacher job satisfaction (*T3JOBSA*) or teacher self-efficacy (*T3SELF*) as the dependent variables (Dumais and Morin, 2019), 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 and years of experience as a teacher. 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. Missing data was handled using listwise deletion, which assumes that the missing data are missing-completely-at-random (Little and Rubin, 2020) which can result in a substantial loss of data and statistical power. The results in Figures II.1.7 and II.1.8 of (OECD, 2020) 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 categorization of the results as positive, no-effect, or negative, provides little information regarding the substantive importance of an 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 (see e.g., Wasserstein and Lazar (2016)), 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 beyond a dichotomous determination of whether or not an effect is statistically significant, and the approach taken in this paper is to compute the probability that the obtained effect is different from zero and to rank countries on the size of those probabilities.[1] 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. Thus, the purpose of this paper is to demonstrate an alternative mode of reporting based on reanalyzing the Figures II.1.7 and II.1.8 from the TALIS report from the perspective of Bayesian statistical inference (see e.g., Gelman et al. (2014); Kaplan (2023)).

The organization of this paper is as follows. In Sect. Overview of TALIS, we provide a brief overview of TALIS 2018. This is followed in Section Preliminaries on Bayesian inference by a review of the key elements of Bayesian statistical inference that are relevant to this paper. A more technical treatment of Bayesian inference is given in Kaplan (2023). In Sect. Analysis of TALIS Dataas a Bayesian HierarchicalModel 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 modeling required for the proper analysis of data arising from complex sampling designs such as TALIS. Then, in Sect. An Example Using TALIS we present the steps of our reanalysis of Figures II.1.7 and II.1.8 in OECD (2020), respectively. This will be followed in Sects. Results for the analysis of teacher job satisfaction and Results for the analysis of teacher self-efficacy 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 Appendix 1: Fig 3, 4, 5, 6. 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 order 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 Appendixes 3 and 4, respectively. Section A proposed Bayesian workflow for ILSA analyses provides a proposed Bayesian workflow that can guide analyses of the type presented in this report, and Section Conclusion concludes with a discussion of the Bayesian advantage as it pertains to the analysis of 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.

Overview of TALIS

In 2008, the Organization for Economic Cooperation and Development (OECD) conducted the first cycle of the *Teaching and Learning International Survey* (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

[1] It should also be emphasized that the frequentist confidence interval provides no additional substantive information about the importance of an effect beyond that of the frequentist p-value.

themes. These 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.

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 paper is based was conducted in 2018. The fourth cycle will be conducted in 2024. Across the cycles of TALIS, the survey design 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 (ISCED2) 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 min 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.

Reporting goals of TALIS

As TALIS is an observational study 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.

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) and Kaplan (2023). 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 modeling 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 (see e.g., Wagenmakers (2007); Wasserstein and Lazar (2016); Kaplan (2023)), such as the methods applied to Figure II.1.7 and therefore can be controversial. We will use the term *frequentist* to describe the paradigm of statistics commonly used today, and which represents the counterpart to the Bayesian paradigm of statistics. Historically, however, Bayesian statistics predates frequentist statistics by about 150 years.

Frequentist probability

Following the discussion given in Kaplan (2023) 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).

Underlying frequentist statistics is the idea of *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 favorable outcomes (here there is only one favorably 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.

This purely theoretical nature of long-run frequency nevertheless 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.

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) who suggested replacing the (infinite) countable additivity axiom with finite additivity and suggested treating probability as *subjective*.[2]

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

The use of the term *subjective* is perhaps unfortunate insofar as it promotes the idea of fuzzy, unscientific, reasoning. Lindley (2007) relates the same concern and prefers the term *personal probability* to *subjective probability*. Howson and Urbach (2006) 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*.

Bayesian inference

The goal of statistical inference is to obtain estimates of the unknown parameters which we denote as θ. For this paper, 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 variation. 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; Laplace, 1774) 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 $p(y|\theta)$ times the prior distribution of the parameters $p(\theta)$ normalized by $p(y)$ so that the posterior distribution sums (or integrates) to one.

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.

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 ahead of time, 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, 2023).

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 1.0. 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 standard deviation of ten, denoted as $\mathcal{N}(0, 10)$, could also be used.

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 stabilizing the estimates of a model, particularly in cases of small sample sizes (see, Gelman (2006)). 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 stabilize the analysis without impacting inferences.

Informative priors

Finally, it may be the case on the basis of previous research, expert opinion, or both, that 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 commits 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 (see e.g. Kaplan (2023); Depaoli et al. (2020), for a discussion of sensitivity to priors).

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; Metropolis et al., 1953). For a nice introduction to the history of MCMC see Robert and Casella (2011).

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 summarize 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) auto-correlation 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. For this paper, 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.

Potential scale reduction factor

When implementing an MCMC algorithm, one of the most important diagnostics is the *potential scale reduction factor* (see, e.g., Gelman and Rubin (1992a); Gelman (1996); Gelman and Rubin (1992b)), often denoted as *Rhat* or $\hat{R}$. This diagnostic is based on analysis of variance and is intended to assess convergence among several parallel chains with varying starting values. Specifically, Gelman and Rubin (1992a) proposed a method where an overestimate and an underestimate of the variance of the target distribution is formed. The overestimate of the variance of the target distribution is measured by the between-chain variance and the underestimate is measured by the within-chain variance (Gelman, 1996). The idea is that if the ratio of these two sources of variance is equal to 1, then this is evidence that the chains have converged. If the $\hat{R} > 1.01$, this may be a cause for concern. Brooks and Gelman (1998) added an adjustment for sampling variability in the variance estimates and also proposed a multivariate extension of the potential scale reduction factor which does not include the sampling variability correction.

The $\hat{R}$ diagnostic is calculated for all chains over all iterations. A problem with $\hat{R}$ originally noted by Gelman et al. (2014) and further discussed in Vehtari et al. (2021) is that it sometimes does not detect non-stationarity, in the sense of the average or variability in the chains changing over the iteration history. A relatively new version of the potential scale reduction factor is available in Stan (Stan Development Team, 2021). This version is referred to as the *Split* $\hat{R}$, and is designed to address the problem that the conventional $\hat{R}$ cannot reliably detect non-stationarity. The *Split* $\hat{R}$ which quantifies the variation of a set of Markov chains initialized from locations points in parameter space. This is accomplished by splitting the chain in two and then calculating the *Split* $\hat{R}$ on twice as many chains. So, if one is using four chains with 5,000 iterations per chain, the *Split* $\hat{R}$ is based on eight chains with 2,500 iterations per chain.

Effective Sample Size

Related to the autocorrelation diagnostic is the *effective sample size* denoted as *n_eff* in the Stan output, which is an estimate of the number of independent draws from the posterior distribution. In other words, it is the number of independent samples with the same estimation power as the T autocorrelated samples. Staying consistent with Stan notation, the n_eff is calculated as

$$\text{n\_eff} = \frac{S}{1 + 2\sum_{s=1}^{\infty} \rho^s} \tag{3}$$

where S is the total number of samples. Because the samples from the posterior distribution are not independent, we expect from Equation (3) that the n_eff will be smaller than the total number of draws. If the ratio of the effective sample size to the total number of draws is close to 1.0, this is evidence that the algorithm has achieved mostly independent draws. Much lower values could be a cause for concern as it signals that the draws are not independent, but it is important to note that this ratio is highly dependent on the choice of MCMC algorithm, number of warmup iterations, and number of post-warmup iterations. As a general rule of thumb, Vehtari et al. (2021) have recommended that the effective sample size be greater than 400.

Summarizing 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.

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 defense of the use of posterior predictive checks can be found in Gelman and Shalizi (2012).

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 mis-specification.

In the Bayesian context, the approach to examining model fit and specification utilizes 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 mis-specification (e.g. omitted variables, incorrect functional form, etc.), poorly specified priors, or both.

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 characterized 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]

Ninety-five percent posterior probability intervals 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 such as pnorm::stats that calculate areas under probability distributions. 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, and this size may be substantively important. This is to be contrasted with the frequentist approach wherein if zero is in the 95% confidence interval, the effect is deemed non-significant (at the 5% level). We present these probabilities in this paper, but it should further be noted that the flexibility available in being able to summarize any aspect of the posterior distribution admits a much greater degree of nuance in the kinds of research questions one may ask. For our paper, we are computing the area between zero and the mean of the posterior distribution, which corresponds to the parameter estimate. We indicate that any interval can be obtained, including the difference between, say 0.10 and the mean of the posterior distribution. Of course, this would render a smaller probability.

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 organizational units (e.g. schools or companies, respectively). Indeed, in many instances, the substantive problem concerns specifically 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

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

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 modeling for the social sciences can be found in Burstein (1980); Goldstein (2011), and Raudenbush and Bryk (2002), among others.

The intercepts and slopes as outcomes model

For this paper, 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 centers on reported job satisfaction of teachers in the United States. Following the TALIS naming conventions, let $T3JOBSA_{ij}$ denote reported job satisfaction of teacher i in school j. We may wish to model $T3JOBSA_{ij}$ as a function of the teacher i took part in induction activities in school j, denoted as $TT3G08_{ij}$. In the empirical example below. The intercepts and slopes as outcomes model can be written as

$$T3JOBSA_{ij} = \beta_{0j} + \beta_{1j}(TT3G08)_{ij} + r_{ij}, \tag{4}$$

where β_{0j} is the intercept that 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) have referred to the model in Equation (4) 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 $\mathbf{z}_j$. 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 reanalyzing, 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

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{5a}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \tag{5b}$$

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 un-modeled between-school variation.

To express Equations (4) and (5a) - (5b) as a Bayesian hierarchical model we specify the following distributions for $T3JOBSA_{ij}$, β_{0j}, β_{1j}, γ_{00}, and γ_{01}. Generally, normal distributions are chosen for regression coefficients because these distributions are conjugate. Conjugate distributions are those that when multiplied by the probability distribution of the data, yield posterior distributions in the same distributional family (Kaplan, 2023). For this model, we specify the following distributions for the regression coefficients,

$$T3JOBSA_{ij} \sim \mathcal{N}[\beta_{0j} + \beta_{1j}(TT3G08)_{ig}, \sigma_j^2], \tag{6a}$$

$$\beta_{0j} \sim \mathcal{N}(\gamma_{00}, \tau_{00}^2), \tag{6b}$$

$$\beta_{1j} \sim \mathcal{N}(\gamma_{10}, \tau_{10}^2), \tag{6c}$$

$$\gamma_{00} \sim \mathcal{N}(\mu_{00}, \omega_{00}^2), \tag{6d}$$

$$\gamma_{10} \sim \mathcal{N}(\mu_{10}, \omega_{10}^2). \tag{6e}$$

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 paper, we choose a non-informative half-Cauchy distribution because it has been shown to be computationally stable as a non-informative prior for variance terms (Gelman, 2006; Kaplan, 2023).

An example using TALIS

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

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 (ISCED Level 2). 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. See the TALIS Technical Report (Dumais and Morin, 2019) for more detail regarding the sampling strata and stratification for the international sampling design and for national sampling designs.

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 (Dumais and Morin, 2019). These weights are calculated as the product of design weights for schools, the design weight for teachers, and the three adjustment factors for teachers and these weights sum to the population of relevant teachers in the country. However, preliminary analyses revealed that more stable convergence of the computing algorithm could be achieved through the use of normalized sampling weights. These normalized 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 normalized weight can be written as

$$NORMWGT_{hij} = \frac{n}{N} * TCHWGT_{hij} \tag{7}$$

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. These normalized weights sum to the number of teachers in the sample for each country.

An important feature of the frequentist analysis of large-scale assessments is the use of balanced repeated replication (BRR) weights that are needed to produce unbiased estimates of sampling error. It should be noted that we did not use BRR weights in this study because, to the best of our knowledge, there is no research examining the use of BRR weights in a Bayesian context, and an investigation into this problem is beyond the scope and purpose of this paper. The development and application of BRR weights in a Bayesian context is an area for future research. For more detail on the construction and use of BRR weights for TALIS 2018, see Dumais and Morin (2019)

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). 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 and also 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 neighbor 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. Predictive mean matching produces unbiased estimates under the assumption that the data are missing completely at random or missing at random (Little and Rubin, 2020). This study assumes that the missing data are missing at random, though we recognize that this assumption may not hold and that the missing data may not be missing at random.

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 paper, we analyzed the first imputed data set via predictive mean matching using the mice package in R (van Buuren, 2012).[4]

[4] We attempted to analyze multiply imputed data sets as per best practice but encountered problems with model convergence. Additional discussion of multiple imputation is presented in Sect. Results for the analysis of teacher self-efficacy, but suffice to say that the analysis of even one multiply imputed data set is better than the use of listwise deletion.

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 easily imputable as 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.

Results for the analysis of teacher job satisfaction

All analyses used the Stan-based software program rstanarm (Goodrich et al., 2020) and all software code used for these analyses are available at https://bmer.wceruw.org/index.html. For these examples, we requested four chains with 5000 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 1000 iterations. Appendices 1 and 2 display the convergence plots for the analysis of Teacher Job Satisfaction and Teacher Self-Efficacy, respectively, from the United States sample only. Focusing on the analysis of Teacher Job Satisfaction, the plots suggest some small concerns regarding convergence, but the *Rhat* and *n_eff* values for the United States shown in Table 1 reveal adequate evidence of convergence.

Our results reveal relatively poor fit to country average teacher job satisfaction with a posterior predictive *p*-value of 0.67. 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 teachers' gender and years of experience as a teacher, ordered in terms of the size of the effect, labeled on the y-axis. The color of the bubble represents ranges of probabilities that the true effect lies between zero and the estimated effect. Of course, bubble plots with other probability ranges can be specified based on what the analyst believes are useful to the substantive question at hand. Nevertheless, 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.

We see in Fig. 1 that countries such as England (United Kingdom) and the UAE showed larger mean estimated effects and relatively large probabilities that the true effect is between zero and the estimated effect. However, note that there are countries with smaller effects that nevertheless have similarly large probabilities that the effect is greater than zero, such as Italy and Austria. Bubble plots are provided for all of the regression analyses included in Figure II.1.7 of OECD (2020) and can be found in Appendix 3.

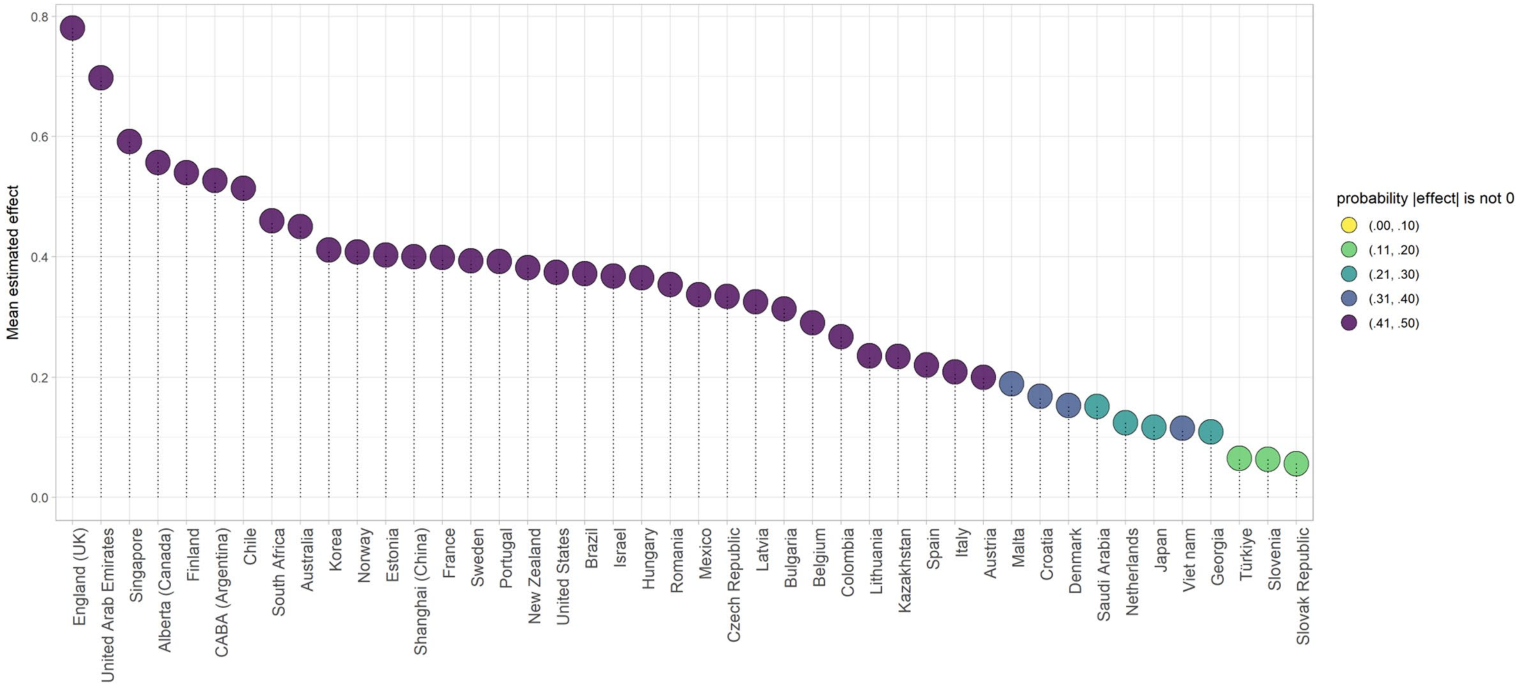


Fig. 1 Bubble plot for reanalysis of Figure II.1.7: Regression of teacher job satisfaction on participation in any induction activities at current school. The *y*-axis is the mean estimated effect, the *x*-axis are the countries, and the color of the bubbles represent ranges of probabilities that the effect is different from zero

Table 1 Participation in any induction activities predicting teacher job satisfaction

Country	**Posterior Mean (sd)**	**95% CI**		**Effective Sample Size**	**Rhat**	**Prob. |effect| ≠ 0**	**Original Results**[a]
		2.5%	**97.5%**				
England (UK)	0.78 (0.21)	0.37	1.21	763	1.0	0.50	**0.84**
United Arab Emirates	0.70 (0.10)	0.51	0.88	1081	1.0	0.50	**0.65**
Singapore	0.59 (0.16)	0.27	0.92	917	1.0	0.50	**0.63**
Finland	0.54 (0.16)	0.23	0.88	947	1.0	0.50	0.48
CABA (Argentina)	0.53 (0.15)	0.23	0.82	955	1.0	0.50	**0.54**
Chile	0.51 (0.19)	0.13	0.89	943	1.0	0.50	**0.44**
Portugal	0.49 (0.13)	0.13	0.64	930	1.0	0.50	**0.44**
Australia	0.45 (0.16)	0.14	0.76	935	1.0	0.50	0.52
Korea	0.41 (0.16)	0.10	0.71	1018	1.0	0.50	0.38
Norway	0.41 (0.11)	0.17	0.62	1016	1.0	0.50	**0.39**
Estonia	0.40 (0.12)	0.17	0.64	971	1.0	0.50	**0.38**
Shanghai (China)	0.40 (0.12)	0.17	0.63	1031	1.0	0.50	**0.43**
Brazil	0.37 (0.13)	0.12	0.65	1072	1.0	0.50	**0.46**
Hungary	0.37 (0.14)	0.10	0.63	1015	1.0	0.50	**0.34**
Romania	0.35 (0.11)	0.14	0.56	1004	1.0	0.50	**0.39**
Mexico	0.34 (0.10)	0.15	0.53	1171	1.0	0.50	**0.29**
Czech Republic	0.33 (0.11)	0.14	0.54	954	1.0	0.50	**0.36**
Latvia	0.33 (0.12)	0.08	0.56	997	1.0	0.50	0.31
Belgium	0.29 (0.10)	0.09	0.49	936	1.0	0.50	0.37
Kazakhstan	0.23 (0.07)	0.10	0.36	1056	1.0	0.50	**0.27**
Alberta (Canada)	0.56 (0.24)	0.10	1.03	1053	1.0	0.49	0.53
South Africa	0.46 (0.19)	0.08	0.85	762	1.0	0.49	**0.79**
France	0.40 (0.17)	0.06	0.73	607	1.0	0.49	**0.43**
Sweden	0.39 (0.18)	0.03	0.73	712	1.0	0.49	**0.38**
Colombia	0.27 (0.12)	0.05	0.51	1074	1.0	0.49	**0.26**
Spain	0.22 (0.09)	0.04	0.40	960	1.0	0.49	**0.31**
New Zealand	0.38 (0.18)	0.02	0.73	1010	1.0	0.48	**0.34**
United States	0.37 (0.18)	0.03	0.72	1048	1.0	0.48	**0.34**
Israel	0.37 (0.17)	0.04	0.69	1056	1.0	0.48	**0.27**
Bulgaria	0.31 (0.15)	0.03	0.60	1021	1.0	0.48	**0.25**
Austria	0.20 (0.12)	-0.03	0.45	1101	1.0	0.45	0.21
Italy	0.21 (0.13)	-0.07	0.69	1101	1.0	0.44	**0.27**
Croatia	0.17 (0.13)	-0.08	0.43	1016	1.0	0.40	**0.21**
Viet Nam	0.12 (0.10)	-0.08	0.31	932	1.0	0.37	**0.09**
Malta	0.19 (0.21)	-0.19	0.64	984	1.0	0.32	0.28
Denmark	0.15 (0.17)	-0.20	0.52	1020	1.0	0.31	**0.27**
Netherlands	0.12 (0,15)	-0.17	0.32	1073	1.0	0.29	**0.13**
Saudi Arabia	0.15 (0.20)	-0.23	0.54	986	1.0	0.28	**0.40**
Japan	0.12 (0.15)	-0.17	0.42	903	1.0	0.28	0.13
Georgia	0.11 (0.18)	-0.22	0.46	990	1.0	0.23	**0.08**
Slovak Republic	0.06 (0.11)	-0.16	0.27	963	1.0	0.19	**0.10**
Slovenia	0.06 (0.14)	-0.20	0.33	1055	1.0	0.18	**0.04**
Lithuania	0.24 (0.10)	0.04	0.44	1094	1.0	0.16	**0.26**
Türkiye	0.07 (0.16)	-0.25	0.38	1030	1.0	0.15	0.13

[a] Statistically significant at 95% confidence level from the original report are indicated in bold

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

An inspection of Table 1 reveals that for most countries, the effective sample size is nearly 1000 indicating low auto-correlation. 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 true effect is between zero and the estimated 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 of the true effect lying between zero and the estimated effect is 0.45. Contrast this with Georgia where zero is in the credible interval but the effect is deemed statistically significant. However, the actual probability that the true effect is between zero and the estimated effect is 0.23. Of course, these interpretations require substantive justification, which would not be possible given the binary (significant/non-significant) decision framework of null hypothesis significance testing.

Results for the analysis of teacher self-efficacy

Appendix 2 displays the convergence diagnostic plots for the analysis of Teacher Self-Efficacy for the United States. As with the analysis of Teacher Job Satisfaction, the analyses of Teacher Self-Efficacy also shows some issues of convergence, however, the *Rhat* and *n_eff* shown in 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 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.

An inspection of Table 2 also reveals that for most countries, the effective sample size is nearly 1000 indicating low auto-correlation. 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 true effect lies between zero and the estimated effect is 0.40. 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 Appendix 4.

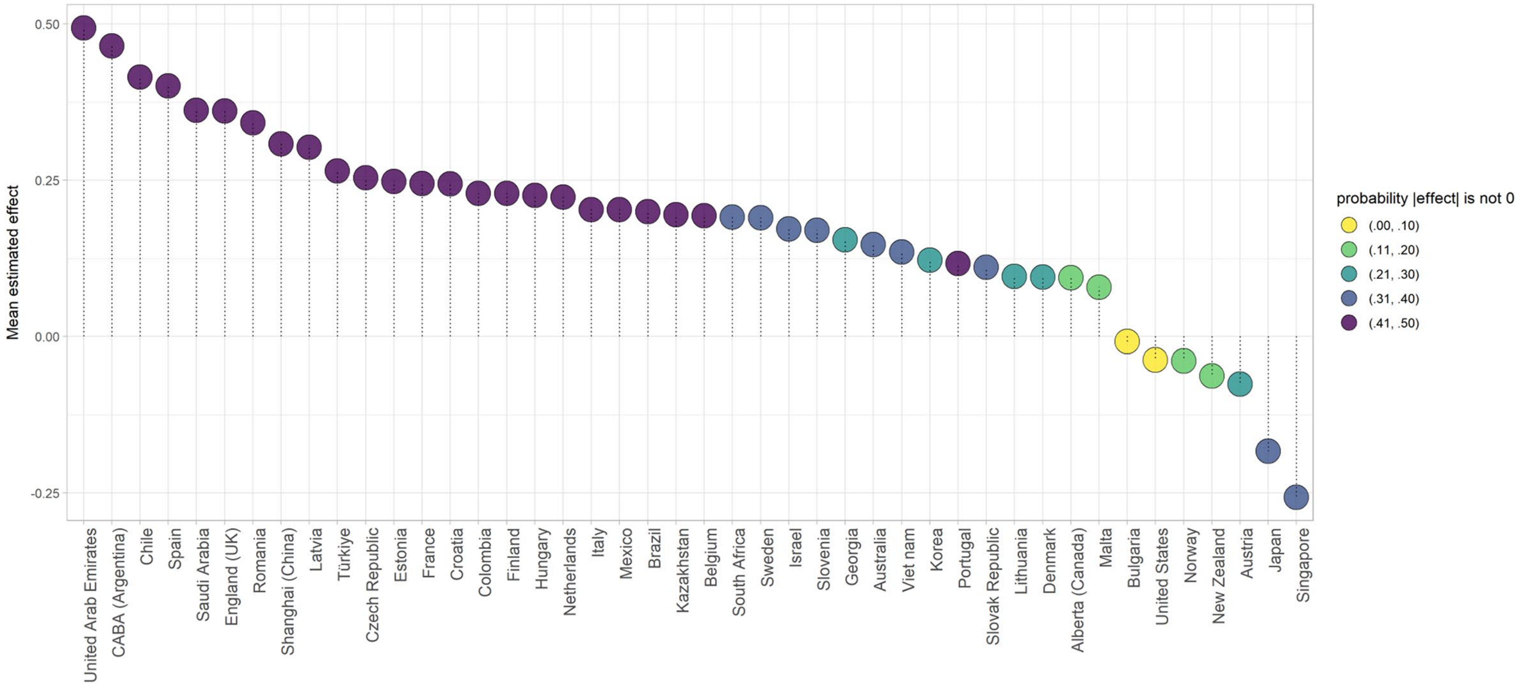


Fig. 2 Bubble plot for reanalysis of Figure II.1.8: Regression of teacher self-efficacy on participation in any induction activities at current school. The *y*-axis is the mean estimated effect, the *x*-axis are the countries, and the color of the bubbles represent ranges of probabilities that the effect is different from zero

Table 2 Participation in any induction activities predicting teacher self-efficacy

Country	Posterior Mean (sd)	95% CI		Effective Sample Size	Rhat	Prob. $\lvert\text{effect}\rvert \neq 0$	Original Results[a]
		2.5%	97.5%				
United Arab Emirates	0.49 (0.09)	0.33	0.66	959	1.0	0.50	**0.48**
CABA (Argentina)	0.47 (0.17)	0.13	0.77	900	1.0	0.50	**0.40**
Spain	0.40 (0.09)	0.24	0.56	794	1.0	0.50	**0.39**
Romania	0.34 (0.13)	0.10	0.59	1007	1.0	0.50	**0.37**
Chile	0.41 (0.18)	0.05	0.75	961	1.0	0.49	**0.39**
Latvia	0.30 (0.12)	0.06	0.55	1047	1.0	0.49	**0.29**
Czech Republic	0.25 (0.09)	0.09	0.43	855	1.0	0.49	**0.27**
Croatia	0.24 (0.10)	0.04	0.45	1063	1.0	0.49	**0.27**
Colombia	0.23 (0.10)	0.02	0.43	1184	1.0	0.49	0.10
Hungary	0.23 (0.10)	0.04	0.43	1006	1.0	0.49	**0.24**
Kazakhstan	0.20 (0.09)	0.02	0.36	941	1.0	0.49	**0.28**
England (UK)	0.36 (0.18)	0.02	0.71	845	1.0	0.48	**0.36**
Estonia	0.25 (0.12)	0.01	0.49	996	1.0	0.48	**0.28**
Belgium	0.19 (0.09)	0.02	0.38	801	1.0	0.48	**0.28**
Saudi Arabia	0.36 (0.19)	-0.01	0.73	1054	1.0	0.47	**0.53**
Shanghai (China)	0.31 (0.16)	0.01	0.62	1060	1.0	0.47	**0.31**
Türkiye	0.27 (0.14)	0.01	0.48	814	1.0	0.47	0.23
Netherlands	0.22 (0.13)	-0.03	0.47	934	1.0	0.46	**0.24**
Italy	0.20 (0.11)	-0.01	0.40	883	1.0	0.47	**0.21**
France	0.25 (0.16)	-0.07	0.56	1069	1.0	0.44	**0.37**
Finland	0.23 (0.14)	-0.03	0.50	1084	1.0	0.46	**0.25**
Mexico	0.20 (0.12)	-0.03	0.44	1000	1.0	0.46	**0.22**
Portugal	0.12 (0.08)	-0.04	0.26	996	1.0	0.44	**0.12**
Brazil	0.20 (0.14)	-0.07	0.47	921	1.0	0.43	**0.18**
Sweden	0.19 (0.15)	-0.11	0.48	1044	1.0	0.40	0.14
Slovenia	0.17 (0.14)	-0.09	0.43	1013	1.0	0.40	0.17
Japan	-0.18 (0.14)	-0.46	0.09	932	1.0	0.40	-0.08
Singapore	-0.26 (0.20)	-0.68	0.14	1090	1.0	0.40	-0.19
Australia	0.15 (0.13)	-0.10	0.41	1057	1.0	0.37	0.13
Viet Nam	0.14 (0.12)	-0.09	0.37	997	1.0	0.37	**0.24**
South Africa	0.19 (0.18)	-0.15	0.56	1071	1.0	0.35	0.01
Israel	0.17 (0.17)	-0.19	0.53	1094	1.0	0.33	0.24
Slovak Republic	0.11 (0.12)	-0.13	0.36	973	1.0	0.32	0.14
Denmark	0.10 (0.12)	-0.13	0.32	838	1.0	0.30	0.08
Lithuania	0.10 (0.12)	-0.13	0.34	917	1.0	0.29	0.13
Korea	0.12 (0.17)	-0.22	0.44	914	1.0	0.27	0.10
Austria	-0.08 (0.11)	-0.29	0.14	1045	1.0	0.26	-0.07
Georgia	0.16 (0.23)	-0.29	0.57	863	1.0	0.25	0.03
Norway	-0.04 (0.08)	-0.18	0.12	954	1.0	0.20	-0.04
Alberta (Canada)	0.09 (0.21)	-0.29	0.51	891	1.0	0.18	0.03
Malta	0.08 (0.20)	-0.32	0.48	946	1.0	0.16	0.21
New Zealand	-0.06 (0.17)	-0.38	0.28	1057	1.0	0.14	0.03
United States	-0.04 (0.15)	-0.32	0.24	1024	1.0	0.10	-0.22
Bulgaria	-0.01 (0.11)	-0.23	0.22	930	1.0	0.03	-0.07

[a] Statistically significant values are indicated in bold

A proposed Bayesian workflow for ILSA analyses

Our analyses of teacher job satisfaction and teacher self-efficacy in Sects. An example using TALIS and Results for the analysis of teacher job satisfaction, respectively, suggest a possible workflow for a Bayesian analysis of large-scale educational data utilizing non-informative or weakly informative priors. Our proposed workflow follows one proposed by (Kaplan (2023), Chapter 12), but of course, other workflows are possible depending on the extent of detail desired in reporting research results (Gelman et al., 2020). Moreover, there are certain similarities between the steps of this workflow and the steps that could be followed in a frequentist analysis of the same data. Fig 3.

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 paper, the outcome variables are composite 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 generalized linear model, but more complex models are, of course, possible. Because this paper is styled to represent the analyses that were conveyed in the original TALIS reports, we utilized linear models, treating each predictor separately. As discussed above, we fully recognize the biases that might occur in treating the predictors separately, but it is beyond the scope of this paper 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 paper.
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) and brms (Bürkner, 2017). It is furthermore critical to appropriately handle missing data. The original TALIS report used listwise deletion, which, as noted earlier, rests on the strict assumption that the missing data are missing-completely-at-random, and can result in a substantial loss of data and statistical power. The state-of-the-art for handling missing data rests on some form of multiple imputation (Rubin, 1987), and for this study, we used multiple imputation under predictive mean matching, though other choices are available. Because we encountered some convergence problems we decided to use the first imputed data set under predictive mean matching. Also, we did not account for the multilevel nature of the data, and admittedly, this could induce some biases in our results. Additionally, it should be pointed out that another legitimate approach to

handling multiply imputed data sets in a Bayesian analysis was proposed by Zhou and Reiter (2010) who recommended analyzing each imputed data set separately and then mixing and summarizing the posterior draws. They find this approach to yield less biased parameter estimates than averaging the parameter estimates.

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. Again, the differences amongst these types of priors is discussed in Sect. Preliminaries on Bayesian inference. 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 stabilize computations, but do not contain very much additional information. Because the goal of the present paper is to mimic the reporting of a policy-relevant report on TALIS, we utilized 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 (2023) and was discussed in Sect. Preliminaries on Bayesian inference. 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 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 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 paper, we provided probabilities the true effect is between zero and the estimated posterior mean.
8. Sensitivity analyses should be conducted, examining the impact of the choice of priors on the substantive results. Other choices of priors can include simple 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 paper, it is likely that results will be robust to reasonable alternative prior distributions.

9. Finally, though it was not discussed in this paper, 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. See Kaplan (2021) for more detail about addressing model uncertainty with examples from large-scale educational assessments.

Conclusion

It is beyond the scope this paper to list all of the advantages of Bayesian methods over frequentist methods. A broader list of advantages can be found in Kaplan (2023) and Wagenmakers et al. (2008), however we list a set of important advantages which have immediate relevance to this paper, and to the analysis and reporting of ILSAs generally.

Summarizing 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; Kaplan, 2021).
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 paper, 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 paper, 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 summarize the entire posterior distribution of the effect - a consequence of treating parame-

ters 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 estimated effect is different from 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.

To conclude, this paper 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) while at the same time demonstrating key differences between the conventional significance testing approach in OECD (2020) 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 recognize that this task is very difficult but we 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 modeling described above.

Appendix 1 Convergence plots for the analysis of teacher job satisfaction: United States

See Figs. 3, 4, 5 and 6.

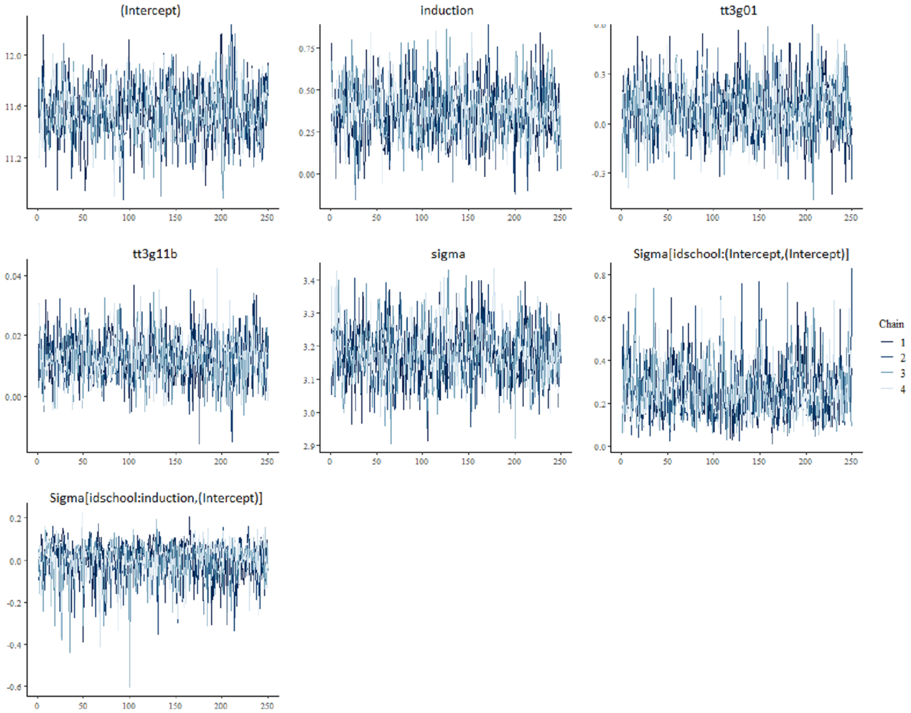


Fig. 3 Trace plots for the model predicting teacher job satisfaction by participation in any induction activities at current school. United States sample. These plots should exhibit a clear rectangular horizontal band over the x-axis. These plots show some problems with the mixing of the chains

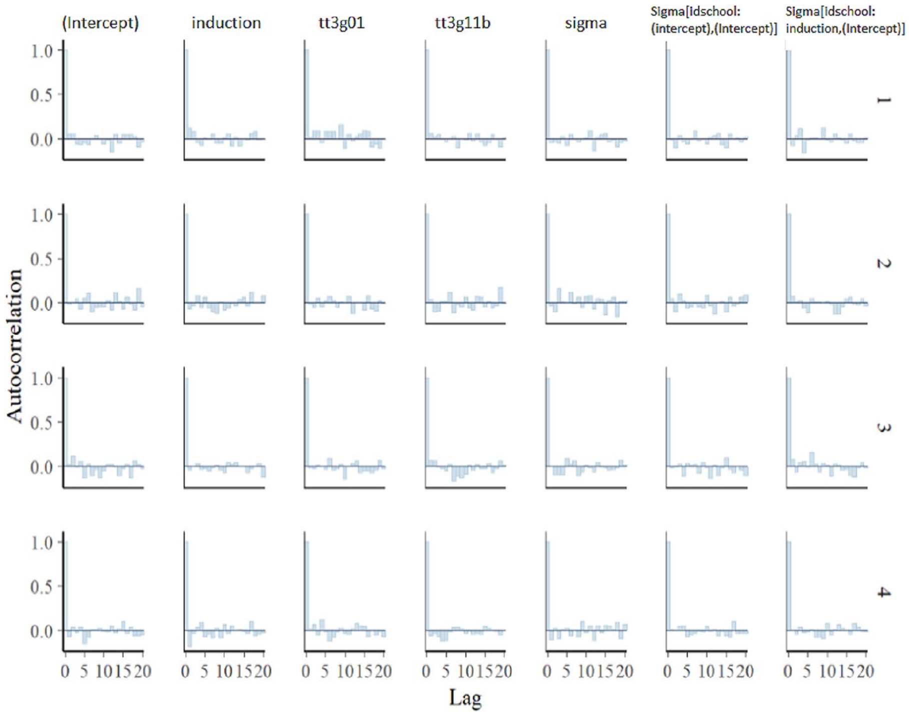


Fig. 4 Autocorrelation plots for the model predicting teacher job satisfaction by participation in any induction activities at current school. United States sample. These plots should show a very high auto-correlation at the first lag and very small auto-correlations thereafter. These plots very low autocorrelation signifying independent draws from the posterior distributions

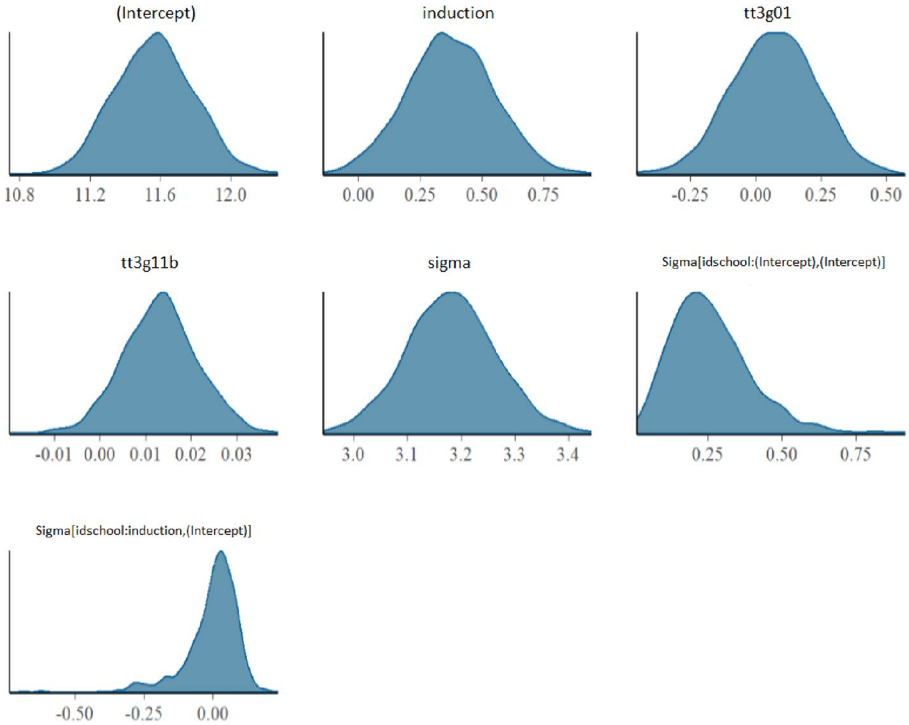


Fig. 5 Posterior probability distribution (density) plots for the model predicting teacher job satisfaction by participation in any induction activities at current school. United States sample. The plots for the regression coefficients and correlations should exhibit more or less a bell-shaped curve while the plot for the variance terms should exhibit a long tail and are bounded below at zero. We note some small problems with the correlation between the error terms of the intercept and slope equations, Sigma[idschool:induction,(Intercept)]

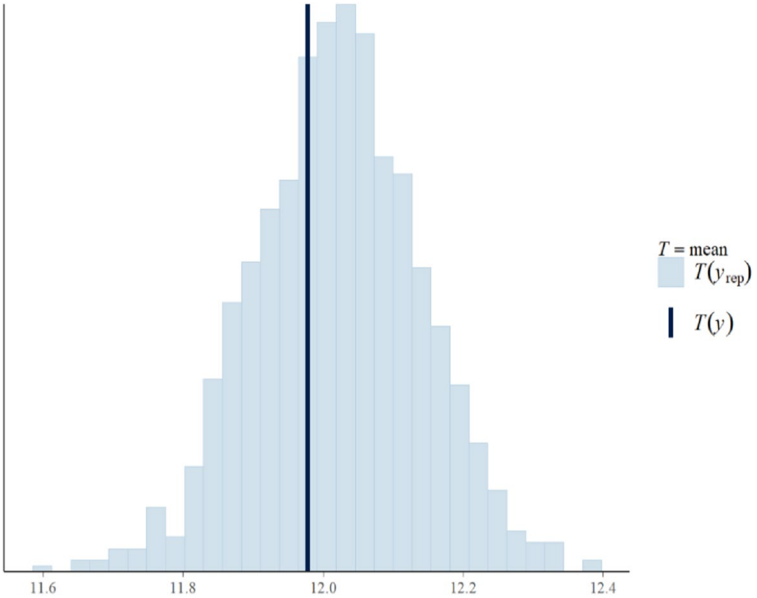


Fig. 6 Posterior predictive check plots for the model predicting teacher job satisfaction by participation in any induction activities at current school ($p = .65$). United States sample. This plot should exhibit a bell-shaped curve with the test-statistic for the data (denoted by the solid black line) positioned at the center of the distribution (0.50), indicating excellent fit to the mean of teacher job satisfaction. We find some misfit to the mean of teacher job satisfaction

Appendix 2 Convergence plots for the analysis of teacher self-efficacy: United States

See Figs. 7, 8, 9 and 10.

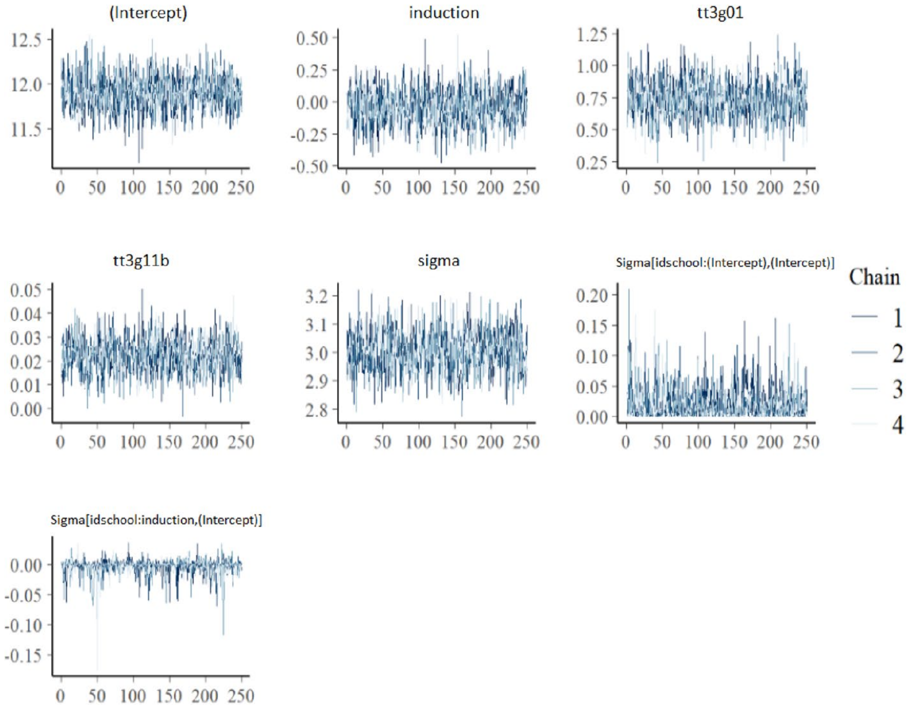


Fig. 7 Trace plots for the model predicting teacher self-efficacy by participation in any induction activities at current school.United States sample. These plots should exhibit a clear rectangular horizontal band over the x-axis. These plots show some problems with the mixing of the chains

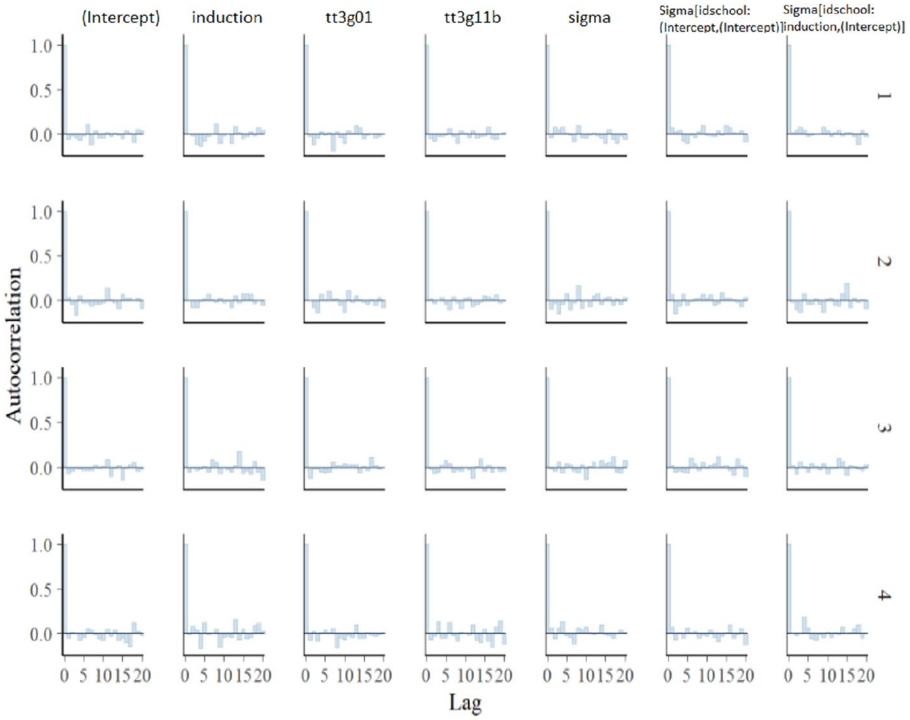


Fig. 8 Autocorrelation plots for the model predicting teacher self-efficacy by participation in any induction activities at current school. United States sample. These plots should show a very high auto-correlation at the first lag and very small auto-correlations thereafter. These plots very low autocorrelation signifying independent draws from the posterior distributions

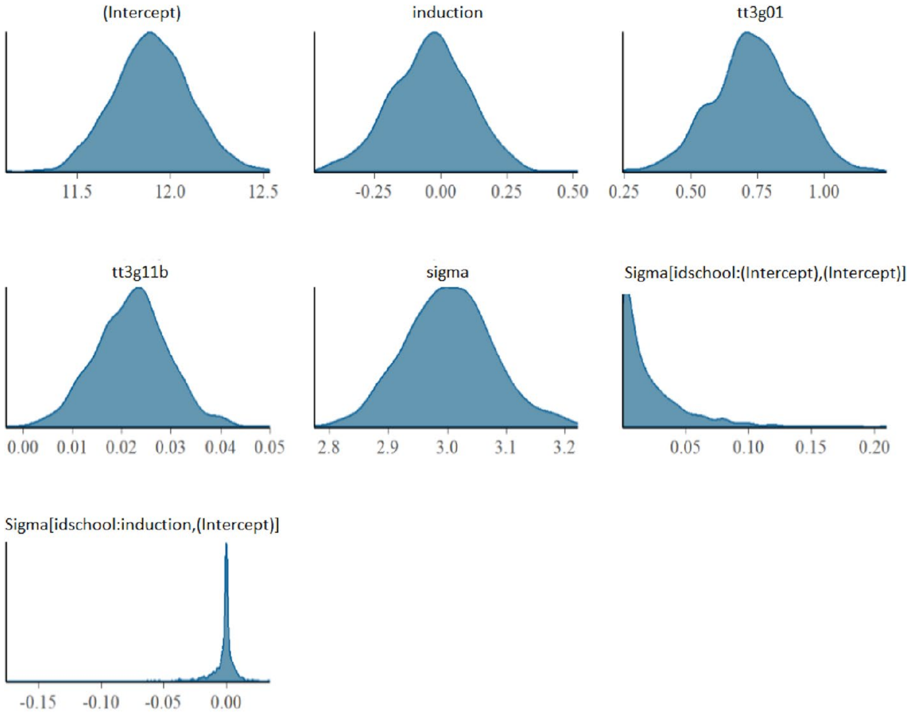


Fig. 9 Posterior probability distribution (density) plots for the model predicting teacher self-efficacy by participation in any induction activities at current school. United States sample. The plots for the regression coefficients and correlations should exhibit more or less a bell-shaped curve while the plot for the variance terms should exhibit a long tail and will be bounded below at zero. The posterior distribution of the effect associated with $t3jobsa_{ij}$ is not as symmetric as desired. This could be improved by increasing the number of MCMC iterations. We also note that the correlation between the error terms of the intercept and slope equations, Sigma[idschool:induction,(Intercept)], is very small

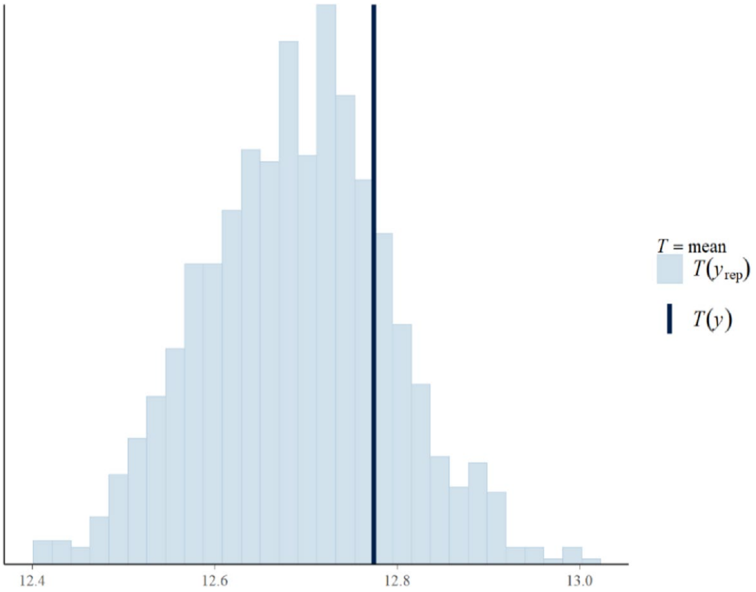
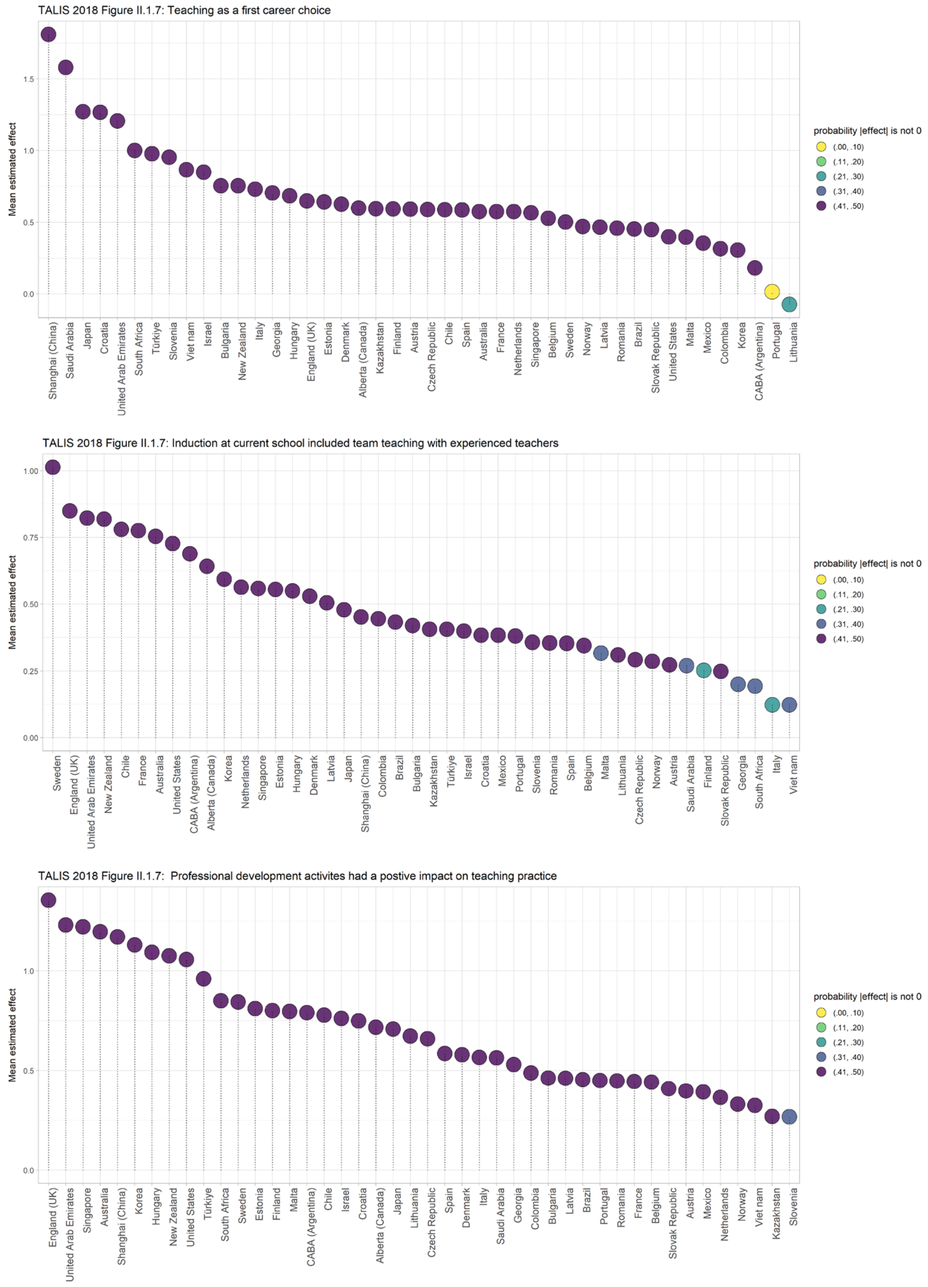


Fig. 10 Posterior predictive check plots for the model predicting teacher self-efficacy by participation in any induction activities at current school ($p = .20$). United States sample. This plot should exhibit a bell-shaped curve with the test-statistic for the data (denoted by the solid black line) positioned at the center of the distribution (0.50), indicating excellent fit to the mean of teacher self-efficacy. We find that the model does not fit the mean very well

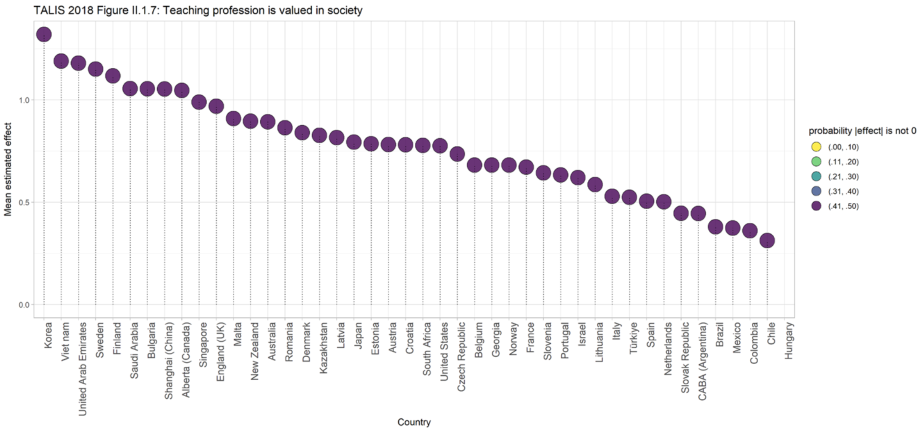
Appendix 3 Results for teacher job satisfaction

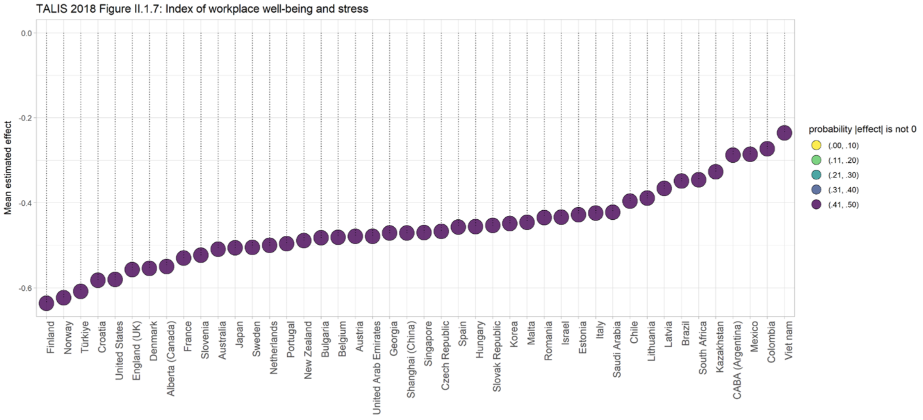
Results for Teacher Job Satisfaction

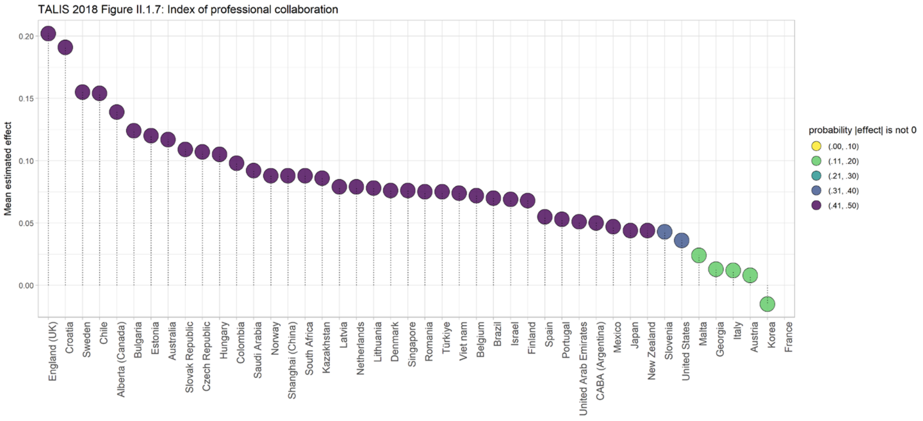

TALIS 2018 Figure II.1.7: Teaching as a first career choice

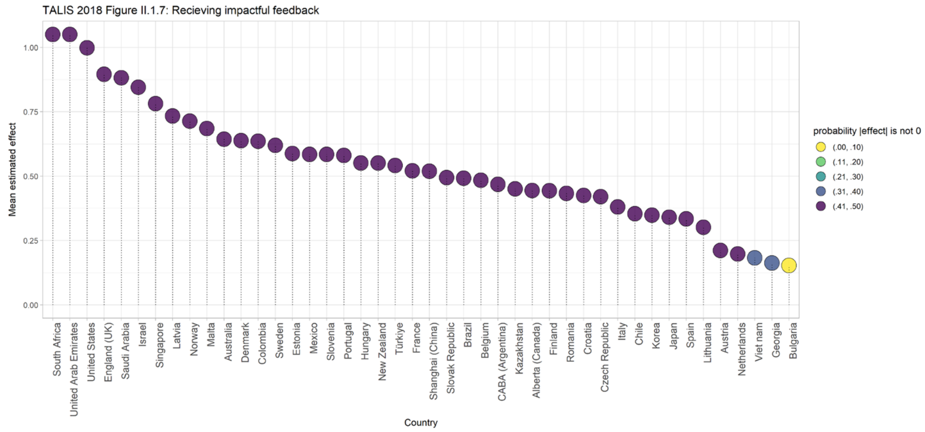
TALIS 2018 Figure II.1.7: Induction at current school included team teaching with experienced teachers

TALIS 2018 Figure II.1.7: Professional development activites had a postive impact on teaching practice

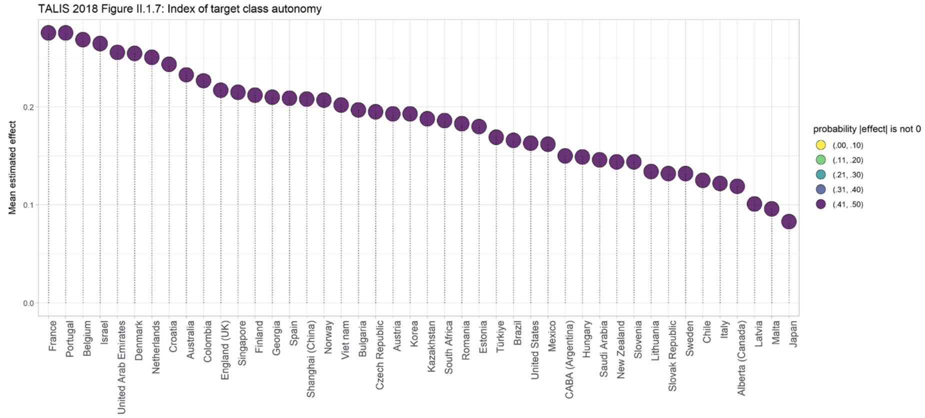

TALIS 2018 Figure II.1.7: Teaching profession is valued in society


TALIS 2018 Figure II.1.7: Index of workplace well-being and stress


TALIS 2018 Figure II.1.7: Index of professional collaboration

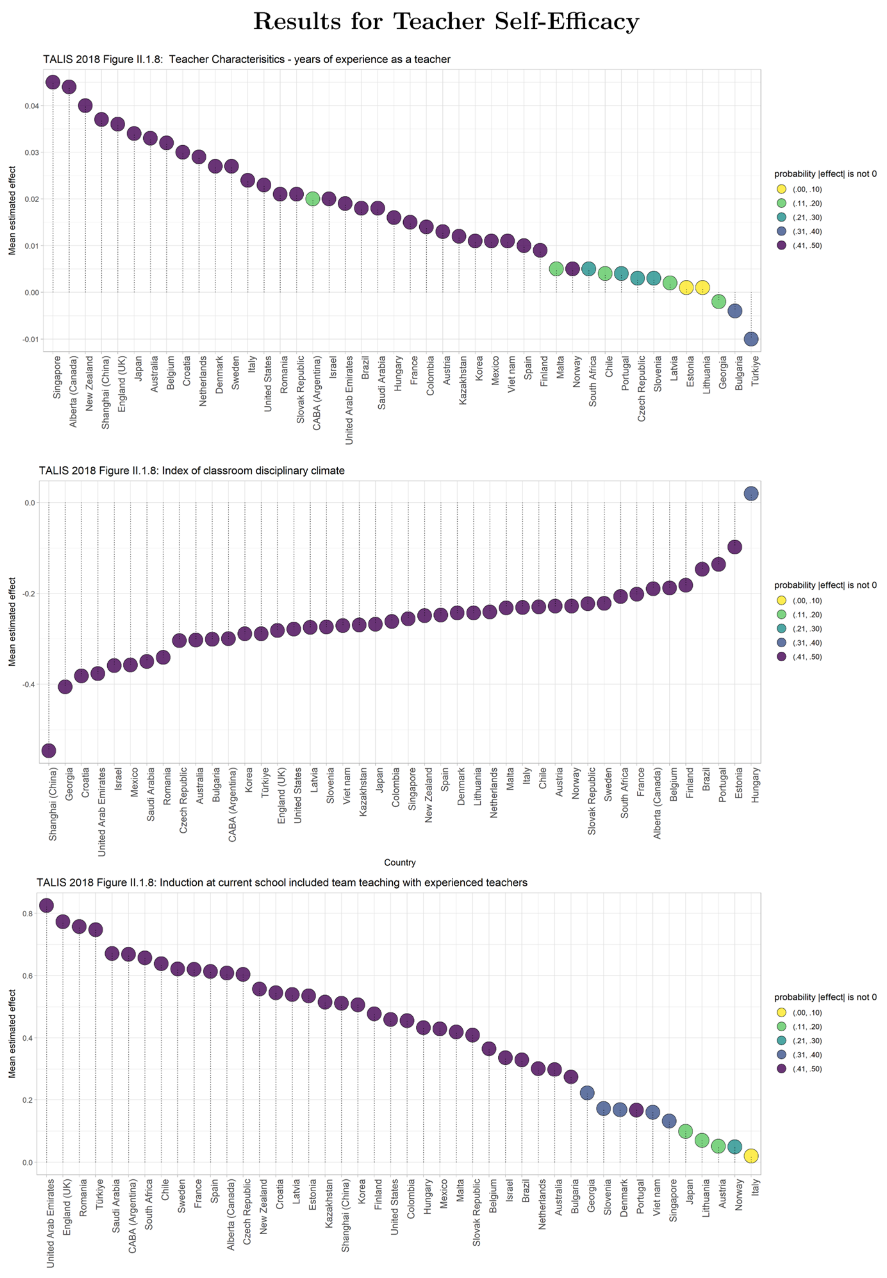


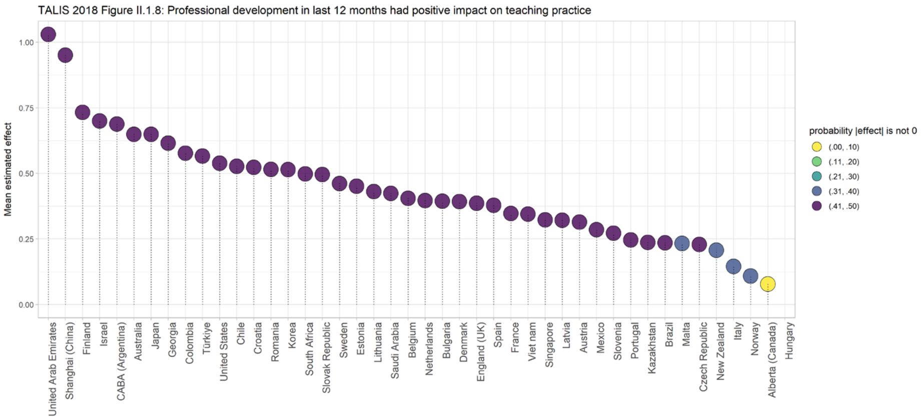
TALIS 2018 Figure II.1.7: Recieving impactful feedback

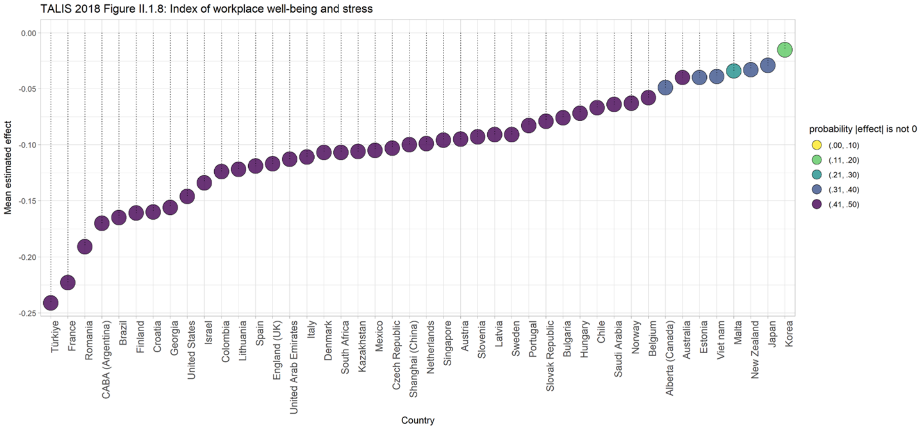


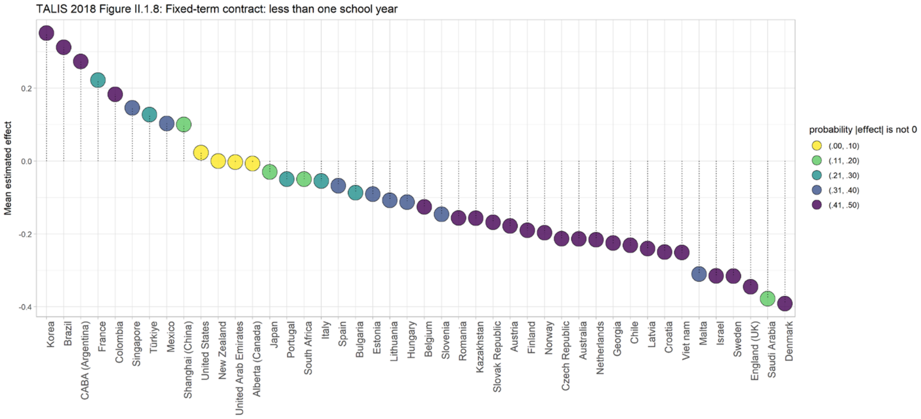
TALIS 2018 Figure II.1.7: Index of target class autonomy

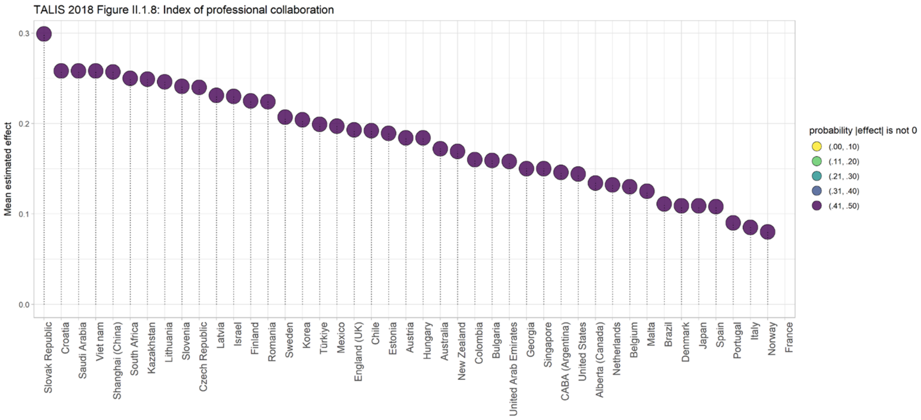
Appendix 4 Results for teacher self-efficacy

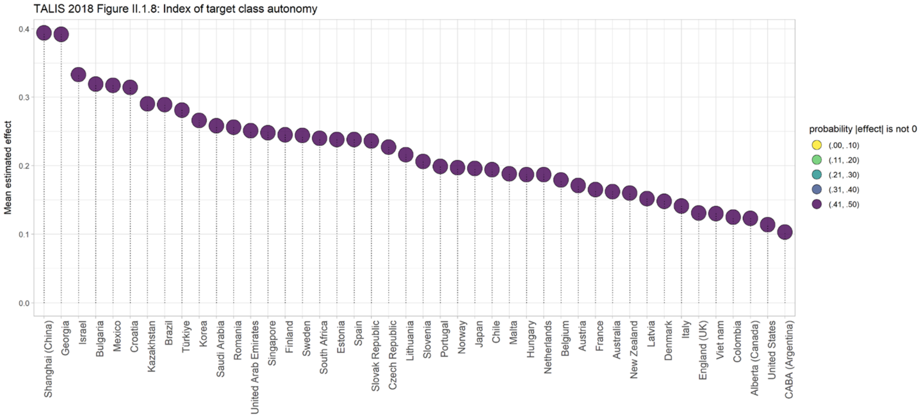



TALIS 2018 Figure II.1.8: Professional development in last 12 months had positive impact on teaching practice
Mean estimated effect
1.00
0.75
0.50
0.25
0.00
United Arab Emirates
Shanghai (China)
Finland
Israel
CABA (Argentina)
Australia
Japan
Georgia
Colombia
Türkiye
United States
Chile
Croatia
Romania
Korea
South Africa
Slovak Republic
Sweden
Estonia
Lithuania
Saudi Arabia
Belgium
Netherlands
Bulgaria
Denmark
England (UK)
Spain
France
Viet nam
Singapore
Latvia
Austria
Mexico
Slovenia
Portugal
Kazakhstan
Brazil
Malta
Czech Republic
New Zealand
Italy
Norway
Alberta (Canada)
Hungary
probability |effect| is not 0
(.00, .10]
(.11, .20]
(.21, .30]
(.31, .40]
(.41, .50]


TALIS 2018 Figure II.1.8: Index of workplace well-being and stress
Mean estimated effect
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
Türkiye
France
Romania
CABA (Argentina)
Brazil
Finland
Croatia
Georgia
United States
Israel
Colombia
Lithuania
Spain
England (UK)
United Arab Emirates
Italy
Denmark
South Africa
Kazakhstan
Mexico
Czech Republic
Shanghai (China)
Netherlands
Singapore
Austria
Slovenia
Latvia
Sweden
Portugal
Slovak Republic
Bulgaria
Hungary
Chile
Saudi Arabia
Norway
Belgium
Alberta (Canada)
Australia
Estonia
Viet nam
Malta
New Zealand
Japan
Korea
Country
probability |effect| is not 0
(.00, .10]
(.11, .20]
(.21, .30]
(.31, .40]
(.41, .50]


TALIS 2018 Figure II.1.8: Fixed-term contract: less than one school year
Mean estimated effect
0.2
0.0
-0.2
-0.4
Korea
Brazil
CABA (Argentina)
France
Colombia
Singapore
Türkiye
Mexico
Shanghai (China)
United States
New Zealand
United Arab Emirates
Alberta (Canada)
Japan
Portugal
South Africa
Italy
Spain
Bulgaria
Estonia
Lithuania
Hungary
Belgium
Slovenia
Romania
Kazakhstan
Slovak Republic
Austria
Finland
Norway
Czech Republic
Australia
Netherlands
Georgia
Chile
Latvia
Croatia
Viet nam
Malta
Israel
Sweden
England (UK)
Saudi Arabia
Denmark
probability |effect| is not 0
(.00, .10]
(.11, .20]
(.21, .30]
(.31, .40]
(.41, .50]





Acknowledgements

An earlier version of this paper was published as OECD Education Working Paper, No. 291. The authors are grateful to Gabor Fülöp, Aliénor Lerouge, Marco Paccagnella and Li Ruochen for valuable comments on the working paper. The comments and suggestions in this paper are solely those of the authors and do not necessarily reflect the views of the OECD.

Author contributions

DK conceptualized the study and guided the design of the study, the statistical analysis, and contributed to drafting the manuscript. KH carried out the analysis as well as contributed to drafting the manuscript. Both authors read and approved the final manuscript.

Funding

This research was supported by contract # EDU/500103553 from the Directorate for Education and Skills, Organization for Economic Cooperation and Development.

Availability of data and materials

The data sets and software supporting the conclusions of this article are available at https://bmer.wceruw.org/index.html.

Declarations

Consent for publication

Not applicable.

Competing interests

The authors declare that there are no competing interests.

Received: 5 April 2023 Accepted: 11 December 2023
Published online: 31 January 2024

References

Bayes, T. (1763). 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, 53 , 370-418. Retrieved from https://doi.org/10.1098/rstl.1763.0053, https://royalsocietypublishin

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434–455.

Bürkner, P.-C. (2017). brms: an R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*, 1–28. https://doi.org/10.1863/jss.v080.i01

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8*, 158–233.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association, 77*, 605–610.

de Finetti, B. (1974). *Theory of probability*. New York: John Wiley and Sons.

Depaoli, S., Winter, S. D., & Visser, M. (2020). The importance of prior sensitivity analysis in Bayesian statistics: demonstrations using an interactive shiny app. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2020.608045

Dumais, J., & Morin, Y. (2019). *TALIS 2018 technical report*. Paris: TALIS, OECD Publishing.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 131–143). New York: Chapman & Hall.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*, 515–533.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehatari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). London: Chapman and Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–511.

Gelman, A., & Rubin, D. B. (1992). A single series from the Gibbs sampler provides a false sense of security. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 625–631). Oxford: Oxford University Press.

Gelman, A., & Shalizi, C. R. (2012). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., & Modrák, M. (2020). Bayesian workflow. *arXiv*. https://doi.org/10.48550/ARXIV.2011.01808

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattn Anal Mach Intel, 6*, 721–741.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). New York: Wiley.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. *R Package Version, 2*(1), 1.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach*. Chicago: Open Court.

Kaplan, D. (2021). On the quantification of model uncertainty: a Bayesian perspective. *Psychometrika, 86*(1), 215–238. https://doi.org/10.1007/s11336-021-09754-5

Kaplan, D. (2023). *Bayesian statistics for the social sciences* (2nd ed.). New York: Guilford.

Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (2nd ed.). New York: Chelsea.

Laplace, P. S. (1774/1951). Essai philosophique sur les probabilities. New York: Dover.

Lindley, D. V. (2007). *Understanding uncertainty*. New York: Wiley.

Little, R. J. A., & Rubin, D. B. (2020). Statistical analysis with missing data (3rd. ed.). New York.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087–1091.

OECD. (2019). TALIS 2018 results (Volume I). Retrieved from https://www.oecd-ilibrary.org/content/publication/1d0bc92a-en, https://doi.org/10.1787/1d0bc92a-en

OECD. (2020). TALIS 2018 results (Volume II). Retrieved from https://www.oecd-ilibrary.org/content/publication/19cf08df-en, https://doi.org/10.1787/19cf08df-en

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Thousands Oaks, CA: Sage Publications.

Robert, C., & Casella, G. (2011). A short history of Markov chain Monte Carlo: subjective recollections from incomplete data. *Statistical Science, 26*, 102–115.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Economic Statistics, 4*, 87–95.

Rubin, D. B. (1987). *Multiple imputation in nonresponse surveys*. Hoboken, NJ: Wiley.

Savage, L. J. (1954). *The foundations of statistics*. New York: John Wiley and Sons Inc.

Stan Development Team. (2021). Stan modeling language users guide and reference manual,version 2.26 [Computer software manual]. Retrieved from https://mc-stan.org (ISBN 3-900051-07-0)

van Buuren, S. (2012). *Flexible imputation of missing data*. New York: Chapman & Hall.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: an improved bR for assessing convergence of MCMC. *Bayesian Analysis*. https://doi.org/10.1214/20-BA1221

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review, 14*, 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York: Springer.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician, 70*, 129–133. https://doi.org/10.1080/00031305.2016.1154108

Zhou, X., & Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *The American Statistician, 64*, 159–163.

Publisher's Note