Educational Review

ISSN: 0013-1911 (Print) 1465-3397 (Online) Journal homepage: https://www.tandfonline.com/loi/cedr20

Bayesian statistics in educational research: a look at the current state of affairs

Christoph König & Rens van de Schoot

To cite this article: Christoph König & Rens van de Schoot (2018) Bayesian statistics in educational research: a look at the current state of affairs, Educational Review, 70:4, 486-509, DOI: 10.1080/00131911.2017.1350636

To link to this article: https://doi.org/10.1080/00131911.2017.1350636

EDUCATIONAL REVIEW, 2018
VOL. 70, NO. 4, 486–509
https://doi.org/10.1080/00131911.2017.1350636

Routledge
Taylor & Francis Group

Bayesian statistics in educational research: a look at the current state of affairs

Christoph König[a] and Rens van de Schoot[b]

[a]Institute of Educational Science, Department of Research Methods in Education, Friedrich Schiller University Jena, Jena, Germany; [b]Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

ABSTRACT

The ability of a scientific discipline to build cumulative knowledge depends on its predominant method of data analysis. A steady accumulation of knowledge requires approaches which allow researchers to consider results from comparable prior research. Bayesian statistics is especially relevant for establishing a cumulative scientific discipline, because the incorporation of background (or prior) knowledge is fundamentally anchored in its basic principles. The aim of the current systematic review is to provide insights into the current state of methodological affairs in educational research, with a focus on Bayesian statistics and the use of prior information. An analysis of publication histories of the 224 educational journals currently listed in the Thomson Reuters Journal Citation Report 2015 indicates that Bayesian statistics is primarily used to solve methodological problems, rather than used to build cumulative knowledge based on a combination of study results with comparable prior research. The utilisation of Bayesian statistics is motivated by its flexibility: models are estimated which would not be estimable with frequentist approaches, thus expanding the methodological repertoire of educational researchers and producing knowledge which otherwise would not have been available. Lastly, the predominant use of noninformative prior distributions indicates that one of the biggest advantages of Bayesian statistics, namely the combination of study results with comparable prior research, remains underutilised in educational research. Practical implications of these findings for educational research are illustrated and discussed.

ARTICLE HISTORY
Received 27 March 2017
Accepted 29 June 2017

KEYWORDS
Bayesian statistics; statistical inference; educational research; educational methods; research methodology; synthesis

Introduction

Quoting Kirk (2003, 100), "The focus of research should be on […] the steady accumulation of knowledge". Cumulative knowledge implies an incremental improvement of estimates of the magnitude and the uncertainty of effects (Kruschke and Liddell 2016). The ability of a scientific discipline to build cumulative knowledge depends primarily on its predominant method of data analysis and interpretation (Schmidt 1996). The reliance of traditional frequentist methods on null-hypothesis significance testing (NHST), which dominates the social

CONTACT Christoph König christoph.koenig@uni-jena.de
Supplemental data for this article can be accessed at https://doi.org/10.1080/00131911.2017.1350636.

sciences, is considered a major hindrance to a steady accumulation of knowledge. A misuse and misinterpretation of its underlying decision rule (if a test statistic is significant, then there is an effect; otherwise there is none) has been identified as the reason for dichotomous thinking in which only significant results matter (Kirk 2003; Schmidt 1996). This type of thinking disregards what is important: (incremental improvement of) the magnitude and the uncertainty of effects and their practical consequences.

Along with increasing criticism of the overreliance on *p*-values, alternatives have been proposed: effect size estimation with confidence intervals with a focus on magnitudes and uncertainties of effects, and meta-analysis and Bayesian analysis for an incremental improvement of these magnitudes and uncertainties (e.g. Cumming 2014; Kruschke, Aguinis, and Joo 2012). The common feature of these alternatives is meta-analytic thinking. Its core premise is that single studies contribute to a given state of research rather than determining it; for a steady accumulation of knowledge, it is necessary to explicitly consider results from comparable prior research (Cumming 2014; Thompson 2002). Combining data at hand with results from comparable previous studies increases the accuracy of parameter estimates and provides the best (i.e. the most complete) available knowledge about a problem (Kruschke, Aguinis, and Joo 2012; Maxwell, Kelley, and Rausch 2008). Meta-analytic thinking allows us to incrementally improve our knowledge of the magnitude and uncertainty of effects. Every study is embedded in a research context, which constitutes the available knowledge about, for instance, the slope of the relation between motivation to learn mathematics and mathematics performance. Bayesian statistics allows this knowledge to be quantified and then combined with new data, resulting in "parameter estimates that are the best available lacking further information" (Kruschke, Aguinis, and Joo 2012, 741). Over time, as the number of similar studies increases, the estimates of the magnitude and the uncertainty of the slope become more precise, making scientific progress truly cumulative.

Effect sizes and meta-analyses already have been a focus of recent reviews of educational research. Sun, Pan, and Wang (2010) show that approximately half of 1243 articles published in 14 journals of the American Educational Research Association (AERA) and American Psychological Association (APA) reported effect sizes, indicating a positive trend. Similarly, Ahn, Ames, and Myers (2012) report an increase in the number of meta-analyses published in AERA and APA journals. Bayesian statistics, despite the considerable increase in attention it has received in other disciplines (e.g. psychology, see van de Schoot et al. 2017; medicine, see Ashby 2006; organisational science, see Kruschke, Aguinis, and Joo 2012), has not. The situation of Bayesian statistics in educational research, thus, remains a blind spot. This is a considerable gap in our knowledge in that Bayesian statistics not only offers inferences richer in information and accuracy, but also has meta-analytic thinking anchored in its basic principles: Bayes' rule allows researchers to explicitly combine their data with results from comparable prior research. The advantage of Bayesian statistics over classical meta-analysis is the possibility to continuously update the available knowledge as new data become available. In meta-analysis, one would have to wait until enough studies are published. Of all alternatives within the meta-analytic framework, Bayesian statistics has the greatest potential for transforming educational science into a truly cumulative scientific discipline. Fully utilising this potential means using appropriate prior information. Bayes' rule requires researchers to include prior information by specifying a prior distribution. This prior distribution can be left uninformative or can be used to integrate previous knowledge. Only this latter option would fully represent meta-analytic thinking and utilise the full potential of Bayesian statistics.

Therefore, it is of great importance that any Bayesian paper describes the role priors play in their paper, and if previous knowledge is indeed integrated into the analyses it should be described in detail (Depaoli and van de Schoot 2017).

The aim of our systematic review is to provide insights into the current state of methodological affairs in educational research, with a focus on Bayesian statistics and the utilisation of background knowledge. With "background knowledge" we refer to the information about model parameters available to researchers prior to observing data, and is meant to distinguish between prior distributions and the knowledge necessary/available to construct this distribution. The research questions are as follows. (1) How is Bayesian statistics used in educational research? (2) What are the reasons to use Bayesian statistics? (3) How do educational studies using Bayesian statistics utilise background knowledge in their analyses? Answers to these questions provide important empirical underpinnings for discussions and information about the current state of education as a cumulative scientific discipline. The structure of this review is as follows. The next two sections illustrate the basic principles of Bayesian statistics and illustrate why Bayesian statistics has the potential for transforming educational science into a truly cumulative scientific discipline. Moreover, it is illustrated and discussed how to combine data at hand with prior information. These two sections are accompanied by a glossary, in which key Bayesian terms used in this paper are introduced and defined (Table 1). In the following the approach to the literature search and review is described and the results of the review are presented. The review is concluded by a discussion of these results with respect to consequences for the research practice within educational science. Highlighted are changes necessary to transform educational science into a truly cumulative scientific discipline.

Statistical inference based on distributions: getting more informative answers to a wider range of questions

Figure 1 shows the posterior distribution of credible values of a regression coefficient (the slope β_1) from an imaginary study investigating the relationship between motivation to learn mathematics and mathematics performance. This distribution is the result of a Bayesian analysis of a simple linear regression model and indicates the plausibility of different values for β_1, conditional on a set of observed data (in this example a random selection of $N = 500$ from the German subsample of PISA 2012). Kruschke (2013b) elegantly points out that Bayesian statistics is simply about reallocating plausibility across a given set of candidate parameter values. The mathematical device for this process is Bayes' rule, the core of Bayesian statistics. Applied to our example Bayes' rule is expressed as follows:

$$p(\beta_1|D) \propto p(\beta_1)p(D|\beta_1)$$

Its key message is that the posterior distribution $p(\beta_1|D)$ is proportional to the product of the prior distribution $p(\beta_1)$ and the likelihood $p(D|\beta_1)$. In our example, a researcher has some background knowledge about the parameter of interest (the slope), which may come from previously conducted analyses or published studies, or even from experts, before the current analysis is conducted. The prior distribution $p(\beta_1)$ summarises this background knowledge as an initial plausibility of different values for β_1, before a data-set is observed. The initial plausibility gets updated once the researcher conducts his analysis and observes the data.

Table 1. Glossary of key Bayesian terms.

Term	Definition
Background Knowledge	A term used to describe the information about model parameters available to researchers prior to observing data. Used to distinguish between prior distribution and the knowledge necessary/available to construct this distribution
Bayes Factor	A quantity summarising the relative support of the data for one model or hypothesis over another. Used to compare competing models or hypotheses
Bayes' Theorem	A formula describing how to update probabilities of model parameters (or hypotheses) when observing data. It centres on conditional probability and its relation to its inverse form
Empirical Bayes	A kind of Bayesian analysis where the hyperparameters of the prior distribution are estimated from the data. Criticised for using the data twice. Contrast to fully Bayesian analysis using prior distributions constructed from background knowledge
Highest Density Interval (HDI)	An interval indicating which part of a distribution is most credible (i.e. is most consistent with the data). It is used to summarise the posterior distribution
Hyperparameters	Parameters governing the shape of a prior distribution. A normal distribution, for instance, is described by its mean and variance
Likelihood	A statistical function of model parameters, assumed to have generated the observed data
Posterior Distribution	A distribution summarising the updated knowledge about model parameters, being a balanced contribution of the prior and the likelihood. Indicates the credibility of different values of model parameters with the data taken into account
Prior Distribution	A statistical distribution summarising all available background knowledge about model parameters prior to observing data. Indicates the credibility of different values of model parameters without any data
Non-informative	A term to describe a prior distribution containing no information about model parameters prior to observing data. It is assumed that no background knowledge is available about the model parameters of interest. Indicated by a large value of the scale hyperparameter (e.g. the prior variance). Typically used in software where priors are automatically built in
Informative	A term to describe a prior distribution containing much information about model parameters prior to observing data. It is assumed that background knowledge based on previous studies is available about the model parameters of interest. Indicated by a small value of the scale hyperparameter (e.g. the variance)
Weakly informative	A term to describe a prior distribution containing some information about model parameters prior to observing data. It is assumed that the only available knowledge about the model parameters of interest is related to its statistical properties. Indicated by a value of the scale hyperparameter reflecting the range of values a parameter can take on (e.g. $max(\beta_1) = SD_y/SD_x$, if x and y are perfectly correlated)
Sensitivity Analysis	An analysis investigating the impact of different specifications of a prior distribution on estimates of model parameters of interest. Used to disclose the robustness of results to different specifications of prior distributions

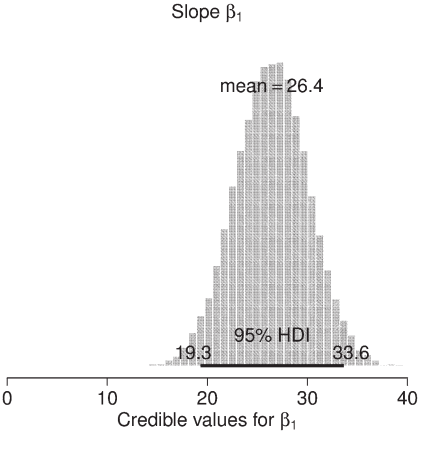

Figure 1. The posterior distribution of credible values of the slope. The most credible value of the slope is shown above the chart (\mu = 26.4). Uncertainty in the parameter estimate is indicated by the 95% HDI, marked as a black bar at the bottom of the chart.

The likelihood $p(D|\beta_1)$ indicates the probability of the data given the model parameter(s); in our example, it is the probability that the data could be generated by a linear regression model with parameter value β_1. The plausibility of different values for β_1 is reallocated based on the information contained in the data, resulting in the posterior distribution $p(\beta_1|D)$, which is a compromise between the background knowledge and the information in the data-set (see Figure 2 for illustrations of the process).

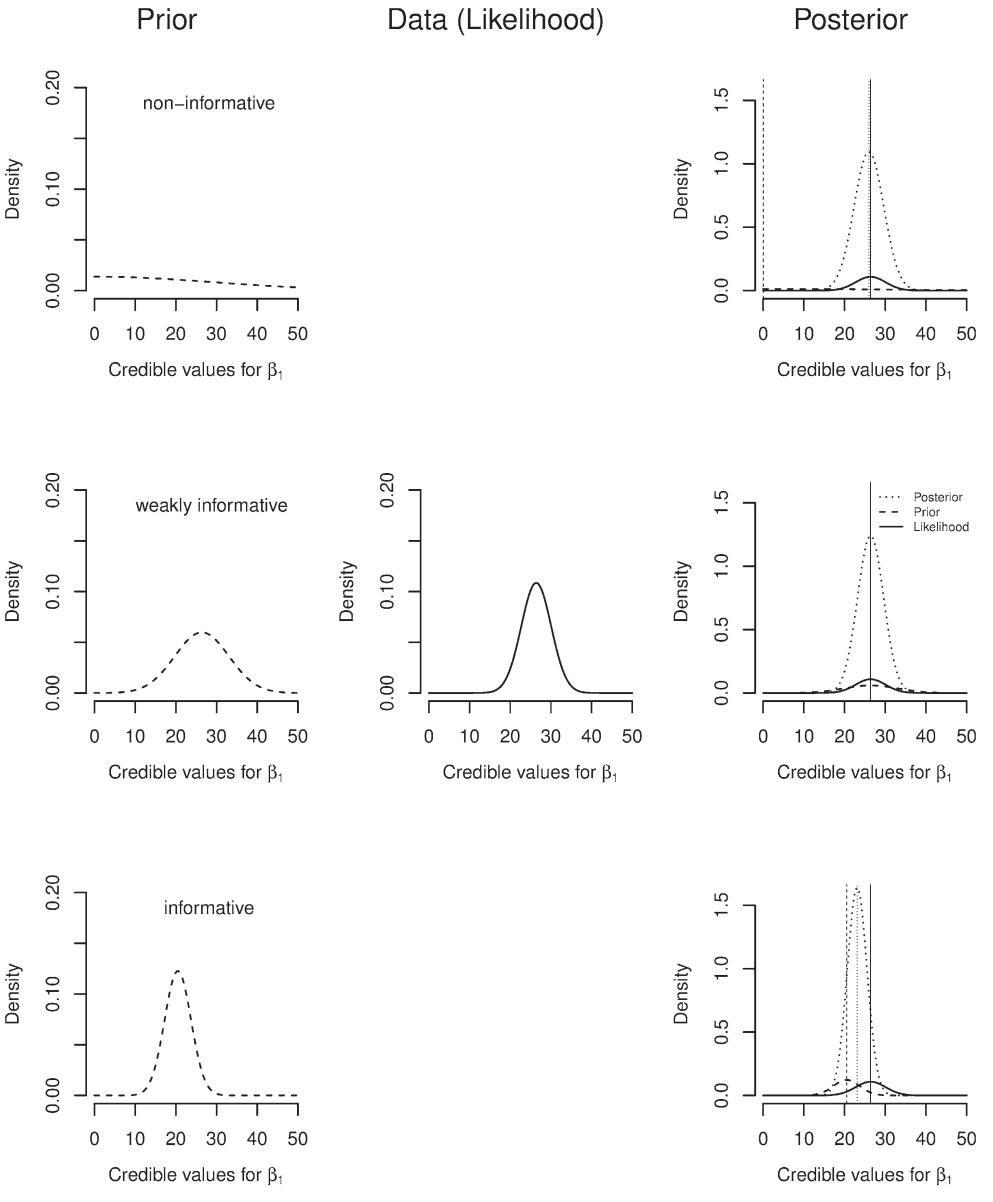

Figure 2. The core principles of Bayesian inference. Information contained in the prior distribution (left side) is combined with information contained in the data (middle), resulting in an updated plausibility of credible values of the slope β_1, summarized in the posterior distribution (right side). The vertical bars in the plots on the right side indicate the means of the respective distributions.

The goal of every data analysis is to yield parameter estimates which are most consistent with the data at hand, to update prior beliefs and to be able to make valid interpretations about a specific research problem. The posterior distribution in Figure 1 indicates which value of the slope is most plausible and consistent with the data at hand. The posterior distribution might be summarised in a point estimate, i.e. the mean (or median, or mode) of the distribution, which is $\mu = 26.4$. Mathematics performance increases by 26.4 points for every 1-point increase in mathematics motivation. The uncertainty of this value is indicated by the width of the distribution. A convenient way to express this uncertainty is the 95% highest density interval (HDI; the black bar at the bottom of the distribution). The 95% HDI indicates that the true value of the increase in mathematics performance likely lies between 19.3 and 33.6 points. In other words, one can be 95% confident that the increase in mathematics performance lies between these two values. This confidence contrasts with classic (frequentist) confidence intervals, which are interpreted as follows: if a similar procedure for constructing a confidence interval were repeated many times using different datasets, then the 95% confidence interval would contain the true parameter value in 95% of the cases. Clearly, the interpretation of the Bayesian 95% HDI is much more straightforward and intuitive.

The potential of Bayesian statistics unfolds with the fact that parameters, such as the slope in a linear regression, are considered random variables which can be summarised by probability distributions (such as the posterior and the prior distributions, as well as the likelihood). In these probability distributions, the degree of uncertainty, for example in the slope, is quantified. They summarise different possibilities within a range of values of the slope estimate. Values that are consistent with the data have a higher probability than values that are inconsistent with the data (Kruschke 2013b). The researcher is now able to answer questions such as: What are the odds that the increase in mathematics performance is higher or lower than a specific value? What is the probability that the increase in mathematics performance exceeds 30 points? What is the probability that the increase in mathematics performance lies between 28 and 35 points? How did the data change the odds that the increase in mathematics performance exceeds 25 points? None of these questions can be addressed in the traditional frequentist framework (Wagenmakers, Morey, and Lee 2016).

Modern Bayesian software allows researchers to flexibly specify complex models, tailored to their questions, which describe their data well. These programs differ primarily in which specifications of a model are up to the user. The commercial software Mplus (Muthén and Muthén 1998–2017) and the open source R-package blavaan (Merkle and Rosseel 2016), for example, do not require the user to specify prior distributions for all model parameters. They work with built-in default prior distributions which facilitate Bayesian data analyses, but may not be appropriate for all models (van de Schoot et al. 2015). Newer open source R-packages, such as JAGS (Plummer 2016) and STAN (Stan Development Team 2017), offer more flexibility in terms of model and prior specification, but at the same time require users to deal with a steeper learning curve. Many package-related websites, however, offer users help and guidance for a plethora of modelling issues, as well as examples of appropriate prior distributions for a wide range of models (e.g. statmodel.com, mc-stan.org, mcmc-jags.sourceforge.net).

Updating knowledge as new data become available: from background knowledge to prior distributions

The prior distribution $p(\beta_1)$ plays a prominent role within Bayes' Theorem. It reflects the amount of available knowledge before observing new data. As mentioned previously, the

prior distribution can be left uninformative. Non-informative prior distributions indicate a lack of background knowledge. Objective Bayesians take the view that non-informative prior distributions, specified per formal rules, should be used when no background knowledge is available, to avoid poor distributions due to systematic elicitation bias: the actual degree of uncertainty about a subject is often underestimated (Berger 2006). Subjective Bayesians, on the other hand, argue that the use of prior information, i.e. informative priors, is necessary and warranted, because it may be the only way to obtain reliable results in absence of large samples (Press 2003). Several studies show beneficial effects of informative prior distributions on the power of small-sample studies (Price 2012; van de Schoot et al. 2015).

Setting up prior distributions is a matter of transforming background-knowledge into adequate distributional forms. They depend on the metric and scale of the parameter of interest. Taking the foregoing example, suppose that, in the year 2021, a researcher is interested in the relation between motivation to learn mathematics and mathematics performance. The parameter of interest is the continuous slope. Its credible values are normally distributed. The shape of the normal distribution is governed by its mean μ and standard deviation σ. While the mean represents the most likely value of the slope, the standard deviation indicates how certain the researcher is about this value. On the left side of Figure 2 the three common kinds of prior distributions are illustrated: non-informative, weakly informative and informative.

In the upper panel (non-informative), the researcher has no background knowledge about the relation and decides to use a prior distribution with a mean of zero and a very large standard deviation, $N(0, 29)$. This distribution assigns higher probabilities to a wide set of possible values of the slope, centred around zero, reflecting how uncertain the researcher is about its value. In the middle panel (weakly informative), the researcher has some knowledge about the relation from a single unpublished study. The researcher decides to use this information and specifies a prior distribution wherein the estimate of the slope and its standard deviation are the prior distribution's parameters, $N(26.4, 6)$. This distribution assigns higher probabilities to a smaller set of possible values of the slope (compared to the previous distribution), centred around the mean of the distribution, which is $\mu = 26.4$. In the lower panel (informative), the researcher knows a meta-analysis of several longitudinal studies on the relation in question. In this case, he has ample knowledge and is quite certain about the value of the slope. He specifies the prior distribution according to the estimate and standard deviation of the slope from the meta-analysis, $N(20.5, 3)$. This distribution assigns the highest probability to a narrow range of possible values centred on $\mu = 20.5$. Due to the large amount of available information, the width of the distribution is very narrow.

The second and third cases represent situations where outcomes of previous studies are used to specify the prior distributions. Such evidence-based prior distributions hold the middle ground between the objective and subjective Bayesian traditions (Kaplan 2014). On the one hand, they are objective priors in a sense that their sources can be verified (van de Schoot et al. 2014). On the other hand, they are subjective priors because they introduce additional information into the analysis, but without the risk to underestimate the uncertainty in this information. Even without results from appropriate previous research, prior distributions almost always contain some information. For example, in the first case the standard deviation of the prior was specified knowing that the maximum value of a regression coefficient is the quotient of the variances of the dependent and independent variables, if they are perfectly correlated. Other examples of such weakly informative prior distributions

are cases when the researcher knows that a value falls between a minimum and a maximum value, or a value that cannot be negative. The specification of prior distributions warrants careful consideration. Detailed descriptions of sources, visualisations of prior distributions and sensitivity analyses to investigate the impact of the prior on the posterior distribution help to increase transparency of and facilitate discussions about the specification of prior distributions (Depaoli and van de Schoot 2017).

While working with non-informative and weakly informative prior distributions can be beneficial, the full potential of Bayesian statistics is utilised only when working with evidence-based prior distributions. In this case, studies are building on each other, thereby incrementally improving the knowledge about the magnitude and uncertainty of parameters: as more information accumulates over time, uncertainty decreases and the researcher becomes increasingly confident about the true value of a parameter. Eventually, this knowledge is the basis for a successful evidence-based educational policy and practice.

The potential of Bayesian statistics for cumulative educational science

To summarise the foregoing illustration, the potential of Bayesian statistics for transforming educational science into a truly cumulative scientific discipline lies firstly in the richer information it provides. In addition to the greater range of possible questions (and associated answers), Bayesian statistics is advantageous because it provides researchers the desired information, namely, the probability of parameter values (or hypothesis) given the observed data at hand. In the traditional frequentist framework, researchers are provided with exactly the opposite information, namely the probability of the data given hypothetical parameter values. Bayesian inference is based only on the data at hand, which is more intuitive and more accurate when it comes to inferring decisions for educational policy and practice. Moreover, they do not rely on auxiliary assumptions for approximating *p*-values or confidence intervals. Unlike the traditional frequentist 95% confidence interval, the 95% HDI truly contains 95% of the most credible values of a parameter.

Secondly, Bayesian statistics offers an alternative approach to inference, based on a combination of prior information and the data at hand which provides the best and most complete knowledge available. Bayesian statistics avoids the "naïve empiricism" of traditional frequentist methods, which describes a mere tallying of supportive (i.e. significant) references (Taleb 2007; in Lambdin 2012). The naïve empiricism is a direct consequence of the black and white decision-making introduced by the sole focus on statistical significance, based on "the [false] belief that if a difference or relation is not statistically significant, then it is zero, or at least so small that it can safely be considered to be zero" (Schmidt 1996, 126). Bayesian statistics allows researchers to constantly update the magnitude and uncertainty of estimates as new data become available. This process is the core of a steady accumulation of knowledge, and the core characteristic of cumulative scientific disciplines.

Literature search and review

To determine the general usage of Bayesian statistics in educational research and the utilisation of background knowledge in educational studies involving a Bayesian data analysis, the literature search and a systematic review of the identified literature included the following stages.

In the first stage, the Thomson Reuters Journal Citation Report, Social Sciences Edition 2014 (Thomson Reuters 2015) was reviewed to identify the educational journals (all types) which are currently listed on that index. In August 2015, using the Web of Science Social Science Citation Index (SSCI) and ProQuest Academic Social Sciences databases, the entire publication history of each of the 224 currently listed educational journals were searched to identify (1) the total number of articles published and (2) the number of articles containing the term *bayes**. This method is common in searches of publication histories in large databases (Mackel and Plucker 2014). The literature search resulted in $N = 265$ articles containing the term *bayes**. The exact search strategy and search terms to replicate the findings can be found in the supplemental material.

In the second stage, these articles were screened to determine the nature of the usage of Bayesian statistics: 57 articles were identified where Bayesian statistics or Bayesian data analysis was mentioned just in one sentence ($N = 28$), mentioned only in the context of the Bayesian Information Criterion (BIC, $N = 5$), just in the references ($N = 18$) or not mentioned at all ($N = 8$). These articles were excluded from further analysis. The screening of the remaining $N = 208$ articles (the full list of references can be obtained from the corresponding author) resulted in the following preliminary categorisation: (1) thematic articles which introduce or discuss Bayesian statistics (or parts thereof) without a statistical empirical application (i.e. commentaries or reviews), (2) methodological articles focusing on comparisons or illustrations of novel or different estimation methods and (3) empirical research articles which addressed a substantial research question with unique samples, i.e. samples which were specifically collected for the respective study. A research assistant was given written instructions and a standardised form for analysing a random subset of the 208 articles. In sum, 41 out of 45 articles (91% agreement) were coded similarly by the first author and the research assistant. Differences regarding the four remaining articles were minor and were resolved by short discussions about the articles. The prevalence of Bayesian statistics in educational research was then calculated by dividing the number of research articles presenting a Bayesian analysis by the total number of published articles. This information was used to answer the first research question.

In the third stage, the methodological and empirical research articles were reviewed with respect to the following aspects: (1) the kind of statistical model/procedure, (2) the authors' motivation to apply Bayesian statistics to a given problem, (3) Bayesian advantages over traditional approaches to a given problem, (4) sample size, (5) choice of priors as stated by the original authors (non-informative, weakly informative, informative), (6) provision of illustrations of the prior distributions and (7) the use of sensitivity analyses for determining the impact of the priors on study results. The first three aspects were used to characterise trends over time in the usage of Bayesian statistics in educational research. The remaining aspects follow recommendations of Depaoli and van de Schoot (2017) to characterise the use of prior information in educational studies involving a Bayesian data analysis. The information was recorded on a standardised form. This information was used to answer the second and third research question.

Results

Overall, among 293,839 articles published in the 224 journals listed by the Thomson Reuters Journal Citation Report, Social Sciences Edition 2014 (Thomson Reuters 2015), Bayesian

statistics was mentioned and/or utilised in 208 articles (0.08%). In 170 of the 224 SSCI educational journals, articles published by approximately three-quarters did not mention the term *bayes**. Thus, Bayesian statistics is present only in about one-quarter of SSCI educational journals. Since the Thomson Reuters Journal Citation Report lists educational journals with the highest impact factor, i.e. the most visible journals, the general visibility of BDA in educational research is, unfortunately, low. Although this is a very small fraction, absolute numbers are steadily increasing since 1960, with a substantial jump in numbers during the turn of the century (1990–2010, see Figures 3 and 4).

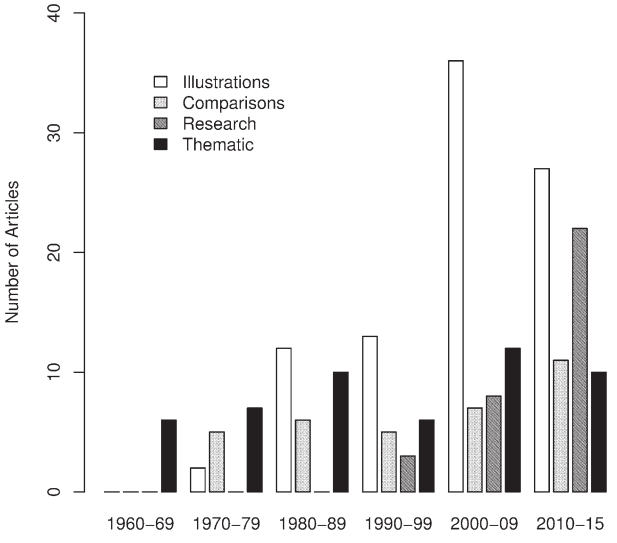

Figure 3. General prevalence of Bayesian statistics in educational research. Absolute frequencies of articles mentioning the term *bayes**

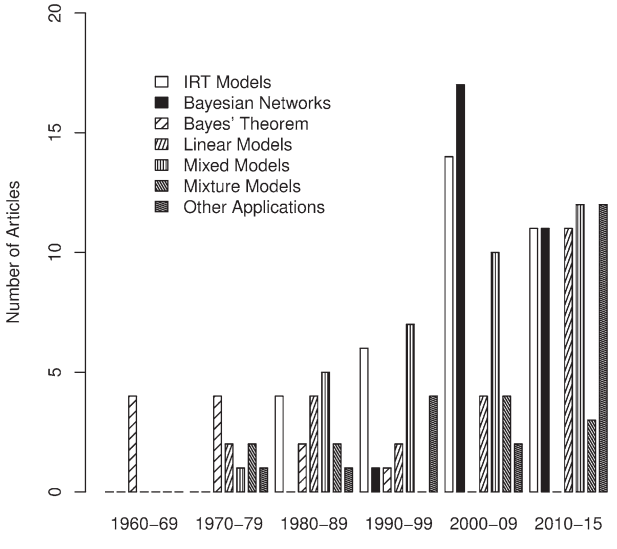

Figure 4. Applications of Bayesian statistics in educational research. Absolute frequencies of different applications of Bayesian statistics according to type of method/model.

Most articles mentioning the term *bayes** were methodological in nature ($N = 127$), including comparisons of Bayesian estimation techniques with other methods ($N = 33$) or illustrations of Bayesian estimation techniques, including simulation studies investigating Bayesian estimation and the development/extension of existing models ($N = 94$). The number of methodological articles grew substantially at the turn of the century (see Figure 3). The next largest category of articles are thematic papers, such as general reviews of statistical methods, book reviews and commentaries ($N = 50$). Thematic papers arguing for a utilisation of Bayesian statistics were present in all decades. While these articles were the majority during the 1960s and 1970s, they are outnumbered by methodological and empirical articles nowadays. Empirical research papers, the last category ($N = 33$), saw an increase in numbers at the turn of the century. In most empirical research papers Bayesian statistics was used to estimate various multilevel, latent-class and ordinary regression models. Only three papers utilised Bayesian statistics for hypothesis testing. Hence, in educational research Bayesian statistics is primarily applied to estimation problems. While Bayesian statistics may be firmly established in the methodological literature, however, its establishment in applied research is lagging. A transfer of knowledge from the methodological to the empirical research literature is only just beginning. The higher numbers of methodological articles, compared to empirical research articles, imply that Bayesian statistics is primarily used to solve methodological problems (e.g. the estimation of complex models), rather than used to answer substantial research questions.

Zooming in on the topics of both the methodological and empirical articles (see Figure 4), Bayesian statistics is used for the estimation and further development of linear, mixed and mixture regression models ($N = 68$, 32.7% of all articles mentioning *bayes**), with a marked increase during the 1980s and 1990s. Another topic for applied Bayesian statistics is the estimation of models based on item response theory (IRT; $N = 36$, 17.3%), where the number of articles also increased substantially during the 1980s and 1990s. Most articles in this category are methodological in nature, i.e. either illustrations or comparisons (primarily with frequentist estimation) of Bayesian estimation of IRT models. It is interesting that, prior to 1992, when Albert (1992) published his article about Bayesian estimation of normal ogive item response curves, Bayesian statistics was only sporadically mentioned in the context of IRT modelling. After Albert's paper, however, the number of articles about Bayesian estimation of IRT models increased substantially. Albert's paper might be considered as a kind of change-point paper, introducing and establishing Bayesian estimation in the context of IRT modelling. While articles with Bayesian statistics applied to both regression and IRT models showed a gradual increase in numbers, the number of articles applying Bayesian Networks ($N = 29$, 13.9%) to computer-based instruction and assessment rocketed during the 1990s. Bayesian Networks are kinds of probabilistic graphical models (PGMs), whose underlying mathematical model is Bayes' Theorem. Each node represents a random variable, and lines connecting them represent probabilistic dependencies. Bayesian Networks are widely used to model uncertain domains, and allow researchers making inferences about the value of a certain node given the observation of values in other nodes in the network (Garcia, Schiaffino, and Amandi 2008). Articles including Bayesian Network analyses are primarily illustrations and evaluations of various computer-based or computer-assisted learning environments. These environments aim at modelling either student performance, student learning or student cognition in general. These papers mention advantages of a Bayesian approach to student modelling only sporadically. It seems that the Bayesian approach to modelling the cognition of students via Bayesian

Networks, and Bayes' Theorem as the underlying mathematical model of the learning environments, is firmly established. There is less need to convince others of the advantages of the Bayesian approach. The direct application of Bayes' Theorem, which was the dominant kind of utilisation of Bayesian principles in educational research prior to 1990 ($N = 10$, 4.8%), has been replaced since then by more sophisticated applications, such as estimation techniques and Bayesian Networks. Other applications of Bayesian statistics ($N = 21$, 10.1%) include Bayesian Hypothesis testing ($N = 6$), the Bayesian way of handling missing data in applied research ($N = 4$), correlational analyses ($N = 4$), Bayesian meta-analysis ($N = 2$) and principal stratification as a means of causal effect estimation ($N = 3$).

In sum, although articles with applications of Bayesian statistics represent only a fraction of all articles published in the 224 journals listed by the Thomson Reuters Journal Citation Report, Bayesian statistics is on the rise in educational research. Moreover, since the 1990s Bayesian statistics is firmly established not only in the methodological literature, but its application in empirical articles has increased and diversified.

Advancing the methodological repertoire of educational research as primary motivation

Motivations to use Bayesian statistics can be summarised by the following three arguments, which are frequently stated in the methodological and empirical research articles: (1) due to inferences being based on distributions, Bayesian statistics provides richer information than frequentist approaches, especially with respect to accounting for uncertainty ($N = 41$); (2) given that Bayesian statistics does not rely on large-sample theory, estimates are accurate in small samples (a motivation frequently stated especially in empirical research articles; $N = 25$); (3) Bayesian statistics allows complex models in combination with complex data structures, which frequentist approaches frequently struggle with, to be estimated with ease and flexibility (especially hierarchical/multilevel models with random effects; $N = 34$). Segawa (2005, 371) summarises this latter advantage nicely: "Furthermore, the hierarchical presentation allows us to utilise a hallmark of the Bayesian MCMC approach: it solves a very complex model […] by successively solving simpler sub-models whose solutions are straightforward and sometimes known".

The general advantages of Bayesian statistics are recognised by researchers (over half of the methodological and empirical research articles mention one of the three aforementioned general advantages: richer inference, small sample behaviour and modelling flexibility). The motivations provided by the authors of the papers imply that Bayesian statistics is used to advance the methodological repertoire of educational research, as well as to produce knowledge which would not have been available with frequentist approaches. More specifically, it is the flexibility of Bayesian inference and the validity of its results in small samples which motivate researchers to utilise Bayesian statistics ($N = 34$). It is interesting to see, however, that the utilisation of prior information is mentioned as a motivation to use Bayesian statistics primarily in methodological articles ($N = 23$). Only one empirical research article explicitly mentions this advantage as a motivation to use Bayesian statistics. Although mentioned as an advantage and motivation to use Bayesian statistics, only four articles use informative prior distributions in their analyses. In the remaining articles, the authors work with non-informative prior distributions. Considering the predominance of non-informative prior distributions in empirical research articles (see below), this implies that utilising background

knowledge is considered rather an impediment than motivation to use Bayesian statistics in empirical research.

Tentative use of background knowledge: a predominance of non-informative priors

The characterisation of the use of background knowledge in empirical research articles utilising Bayesian statistics follows the recommendations of Depaoli and van de Schoot (2017). This section is structured accordingly. The studies included in this part of the review (empirical research articles as identified in the foregoing stages of the literature search) are summarised in Table 2.

Choice of prior and justification. In a large proportion (38.71%) of empirical research articles either program-specific prior distributions are used or the data are used to specify the prior distributions for the respective analyses. Estimating parameters of prior distributions alleviates the problem of prior specification to a certain degree. This practice, however, is criticised for its data double-dipping (Darnieder 2011). Among articles implementing a Fully Bayes approach, except for Gilger, Pennington and Defries (1991), Bekele and McPherson (2011) and Fraile and Bosch-Morell (2015), non-informative or weakly informative prior distributions are predominant. Weakly informative prior distributions are used by Buckley and Schneider (2005) and Gudmestad, House and Geeslin (2013) for very specific parameters: the most-likely class membership in a latent class model and a covariance matrix. These two parameters are good examples of situations where a subjective approach (i.e. the elicitation of informative priors) might not be feasible and that instead warrant the use of non-informative or weakly informative prior distributions. There is seldom background knowledge available about specific parameters such as most-likely class membership or the elements of covariance matrices. Very few authors explicate the use of prior distributions. Buckley and Schneider (2005), Doyle (2010), Doyle and Gorbunov (2011) and Gudmestad, House and Geeslin (2013) state that no background knowledge about the parameter(s) of interest was available. The studies by Karpudewan, Ismail and Roth (2012), Karpudewan, Roth and Ismail (2015) and Li and Shen (2013), all of which conducted Bayesian *t*-tests and hypothesis testing based on Bayes factors, contained no information about the prior distributions utilised in their analyses. While posterior distributions for continuous parameter estimates typically are robust against moderate changes in the vagueness of broad priors, Bayes factors are extremely sensitive to the choice of prior and it is important to use meaningfully informed prior distributions (Gelman and Shalizi 2013). Without meaningfully informed priors, Bayes factors can be quantitatively inaccurate and/or meaningless (Vanpaemel 2010; Kruschke 2013a).

Sources of priors. Four empirical research papers do not provide any information about the prior distributions used in their analyses. It is likely that they used the default, non-informative prior distributions available in the respective statistical programmes. The weakly informative priors used by Buckley and Schneider (2005) and Gudmestad, House and Geeslin (2013) are specified in accordance with recommendations from methodological articles. These articles provide either full model specifications or specifications of prior distributions. These prior distributions, however, are specified and used for statistical reasons (e.g. to facilitate model estimation) rather than for incorporating substantial background knowledge into the analyses. Gilger, Pennington and Defries (1991) specify their priors according to

Table 2. Characteristics of the use of background knowledge in empirical research articles.

Author(s)	Application of Bayesian Statistics	Sample Size (*N*)	Choice of Priors	Source of Priors
Gilger, Pennington, and Defries (1991)	Application of Bayes' Theorem	1555 individuals	Informative Reason: replication	Prior studies
Zwick (1993)	Linear Model	2624–3057 individuals	Data-dependent (Empirical Bayes)	–
Seltzer, Frank, and Bryk (1994)	Mixed Model	2500 individuals	Data-dependent (Empirical Bayes)	–
Heck (2000)	Mixed Model (Supplemental Analysis)	6970 in 122 schools	Data-dependent (Empirical Bayes)	–
Tobias (2002)	Linear Model	160–381 individuals	Non-informative No reason stated	–
Buckley and Schneider (2005)	Mixture Model	37–201 individuals	Non-informative Weakly informative Reason: methodological literature	Specification according to methodological literature
May and Supovitz (2006)	Mixed Model	55,932–56,693 individuals	– [a]	–
Meyer and Xu (2007)	Bayesian Network	16,914 individuals	Informative (prior probabilities) Reason not applicable (software-specific)	–
Notenboom and Reitsma (2007)	Mixture Model	458 individuals	–	–
Rowan and Miller (2007)	Mixed Model (Supplemental Analysis)	830 leaders, 5533 teachers in 114 schools	Data-dependent (Empirical Bayes)	–
Jin and Rubin (2009)	Principal Stratification	1090 individuals	Non-informative No reason stated	–
Doyle (2010)	Linear Model	115 individuals	Non-informative Reason: no prior info available	–
Petscher (2010)	Meta-Analysis (Mixed Model)	32 studies	Data-dependent (Empirical Bayes) HLM program default	–
Bekele and McPherson (2011)	Bayesian Network	571 individuals	Informative (prior probabilities) Reason not applicable (software-specific)	–
Clewley, Chen, and Liu (2011)	Bayesian Network	65 individuals	Informative (prior probabilities) Reason not applicable (software-specific)	–
Doyle and Gorbunov (2011)	Mixed Model	50 individuals	Non-informative Reason: no prior info available	–
Scarpino et al. (2011)	Mixed Model	85 individuals	Non-informative No reason stated	–
Boyd et al. (2012)	Mixed Model (Supplemental Analysis)	65,000–80,000 individuals	Data-dependent (Empirical Bayes)	–
Galbraith and Merrill (2012)	Mixture Model	1675 scores of 112 courses	Non-informative (software-specific)	–
Galbraith, Merrill, and Kline (2012)	Mixture Model	116 courses	Non-informative (software-specific)	–
Hoogerheide, Block, and Thurik (2012)	Linear Model	8244 individuals	Non-informative No reason stated	–
Karpudewan, Ismail, and Roth (2012)	Bayesian *t*-test (Hypothesis Testing)	263 individuals	–	–
Laru, Näykki, and Järvelä (2012)	Bayesian Network	21 individuals	Informative (prior probabilities) Reason not applicable (software-specific)	–

Ageyi-Baffour et al. (2013)	Linear Model	238 individuals	–	–
Li and Shen (2013)	Bayesian *t*-test (Hypothesis Testing)	48 individuals	–	–
Tammets, Pata, and Laanpere (2013)	Bayesian Network	16 individuals	Informative (prior probabilities) Reason not applicable (software-specific)	–
Wheadon (2013)	Bayesian IRT	267 individuals in 11 schools	Non-informative (software-specific; WinBUGS)	–
Gudmestad, House, and Geeslin (2013)	Linear Model	6342 individuals	Weakly informative Reason: methodological literature	Specification according to methodological literature
Suchodoletz and Gunzenhauser (2013)	Linear Model (Path Analysis)	60 Individuals	–	–
McDermott et al. (2014)	Bayesian IRT (Supplemental Analysis)	3077 individuals	Data-dependent	–
Cummings et al. (2015)	Mixed Model	637079 individuals in 8967 schools	Data-dependent (Empirical Bayes)	–
Fraile and Bosch-Morell (2015)	Mixed Model	2410–13662 questionnaires of 670–726 individuals	Informative Reason: longitudinal study	Estimates of analysis of year 1 data
Karpudewan, Roth, and Ismail (2015)	Bayesian *t*-test (Hypothesis Testing)	67 individuals	–	–

results of prior studies. Fraile and Bosch-Morell (2015), in their 2-year study of lecturer evaluations, use the results from year one and specify the prior distribution for their subsequent analysis accordingly.

Illustration of priors and sensitivity analyses. Except for Buckley and Schneider (2005) and Fraile and Bosch-Morell (2015), none of the studies included in this review contain illustrations of the priors used in the analyses. Although such illustrations are of no direct value to the primary analyses in the studies, they would nevertheless facilitate the understanding of the prior distributions. Similarly, none of the studies present sensitivity analyses for investigating the impact of their (non-informative) prior distributions on the results of the primary analysis.

The relatively large proportion of empirical research articles utilising the Empirical Bayes approach and non-informative prior distributions (whose uses are generally left unexplained) implies a very tentative utilisation of background knowledge. Bayesian statistics in educational research is, consequently, firmly established in the objective Bayesian tradition, and less in either the subjective or evidence-based Bayesian tradition. It is worth mentioning, however, that the number of empirical research articles implementing the Empirical Bayes approach, where parameters of prior distributions are estimated from the data, is growing slower relatively to the number of articles implementing a Fully Bayes approach, where the specification of prior distributions is required (from the 1980s until the 2010s N_{EB} increased from four to 11 articles, while N_{FB} increased from four to 33 articles). The predominant use of non-informative prior distributions may be interpreted as an attempt to stay on familiar ground or, in other words, the utilisation of Bayesian statistics is considered as a simple alternative to frequentist estimation methods. A representative example for trying to stay on familiar ground is the study by Suchodoletz and Gunzenhauser (2013). The authors explicitly state that the primary motive for utilising Bayesian statistics was to maximise the accuracy of parameter estimates in case of small samples (the sample sizes in their study range from $N = 60$ to $N = 201$). They did not report the exact specification of their prior distributions and did not conduct any sensitivity analyses. Sensitivity analyses are important in small sample studies, because effects of non-informative prior distributions on small sample analyses are still unclear (Depaoli and van de Schoot 2017). Moreover, although they followed a Bayesian approach to data analysis, they still reported *p*-values. This practice is illustrative for the attempt "to make the Bayesian omelette without breaking the Bayesian egg" (Savage 1962). The transition towards the full utilisation of Bayesian statistics in educational research is still in its infancy.

Discussion

The aim of this review was to provide insights into the current state of methodological affairs in educational research, with an explicit focus on applications of Bayesian statistics and the utilisation of background knowledge. It was found that since the 1990s Bayesian statistics is not only firmly established in the methodological literature, but its application in empirical articles has increased and diversified. The primary motivations to use Bayesian statistics are an increased accuracy of parameter estimates (especially in small samples), richer inference and results more meaningful for educational practice and the possibility to estimate complex models which frequentist estimation methods often struggle with. The transition towards the full utilisation of Bayesian statistics in educational research, however, is only just

beginning. Studies utilising background knowledge through the specification and use of informative prior distributions constitute only a very small fraction of all articles utilising Bayesian statistics. Currently, Bayesian statistics is recognised and used simply as a powerful alternative to frequentist estimation methods. The full potential of Bayesian statistics, especially with respect to the accumulation of knowledge and for education as a truly cumulative scientific discipline, is yet to be explored.

The situation is comparable with that of the organisational sciences, where fewer than half of one percent of articles published between 2001 and 2010 applied Bayesian statistics (Kruschke, Aguinis, and Joo 2012). The trend in the prevalence of Bayesian statistics in educational research is remarkably similar to that in psychology. Following a 10-year delay compared to the general statistical literature (Andrews and Baguley 2013), both psychological and educational research increased their coverage and use of Bayesian methods in the 2000s. Reasons for this relatively slow adoption of Bayesian methods are frequently discussed and include the statistics training of students relying predominantly on traditional statistical methods (Henson, Hull, and Williams 2010) and the relative ease with which traditional analyses can be conducted (Kaplan and Depaoli 2013). Regarding the latter aspect, we face some kind of statistical catch-22 (in the sense of van de Schoot et al. 2017). The principles of Bayesian statistics are easy to learn. To fully utilise its potential, however, it is necessary to delve into more complex modelling issues. While the standard models are well covered by accessible software such as Mplus, successfully handling more complex situations involving, for instance, multilevel or longitudinal models, requires researchers to embark with more complicated software, which might deter interested novice Bayesian researchers. An increasing number of textbooks, however, aim at attenuating the learning curve (e.g. Kaplan 2014; Kruschke 2015).

The consensus appears to be that Bayesian methods are not only a novel set of tools for data analysis but also require a (far-reaching) rethinking of the way scientific knowledge is created, updated and accumulated. The hesitant use of background knowledge in the studies included in this review, reflected by the predominance of non-informative priors, indicates that this rethinking is still in its infancy. Although certain useful advantages of Bayesian statistics are recognised, the full epistemological shift has not yet occurred.

Specifying informative prior distributions: difficulties and possibilities

To fully embrace meta-analytic thinking, Cumming (2014, 23) states that "any one study […] needs to be considered alongside any comparable past studies and with the assumption that future studies will build on its contribution". Bayesian statistics enables researchers to do exactly that. Although the quantification of background knowledge is the most striking advantage of Bayesian statistics, it is also arguably its most controversial aspect. From the frequentist viewpoint, the use of prior information is criticised for being inherently subjective. Both Bayesian and frequentist inference, however, contain subjective elements. Kruschke (2013b) illustrates, for example, how *p*-values and their interpretations change depending on researchers' implicit assumptions and intentions. The important difference is that prior distributions are explicit assumptions and testable parts of a statistical model; they can be discussed and inspected in terms of their impact on results, whereas researchers' intentions cannot (Gelman and Shalizi 2013; Kadane 2011; Kruschke, Aguinis, and Joo 2012).

It must be acknowledged that specifying informative prior distributions can be very difficult. These difficulties can be illustrated with the following three aspects. First, the selection of previous research to inform the prior admittedly remains subjective. Moreover, just like meta-analysis, even the evidence-based subjective approach for the specification of prior distributions (Kaplan 2014) suffers from publication bias and the file drawer problem. It is nevertheless an explicit choice that opens room for fruitful and necessary discussions within the discipline (Kadane 2011).

Second, it must be determined to what extent conditions of previous research apply to a new study. It can be challenging to decide which studies "count" as valid background knowledge to construct prior distributions from. Questions arise concerning the comparability of studies conducted in different contexts, with different samples and different variables. Using studies from different contexts might imply an unwarranted generalisation; excluding studies based on their contexts might be too restrictive and imply that no background knowledge is available, when in truth there is. Using studies with different variables (e.g. in structural equation models) may be criticised as well, although in this case it can be argued that it is just another missing data problem which can be modelled accordingly. Differences in samples, procedures and covariates introduce uncertainties which have to be taken into account when specifying prior distributions according to such previous research. In this regard, power priors are the topic of recent discussions. These kinds of prior distributions allow researchers to attach weights to different sources of background knowledge, based on their similarity to the research problem at hand (e.g. Ibrahim et al. 2015). It is also possible to conduct a mixed-effects meta-analysis on previous research, which accounts for study heterogeneity in the outcomes (Kirkham, Riley, and Williamson 2012). The results of such meta-analyses can then be used as parameters for the prior distribution in the current analysis. There are no easy answers, and the field of Educational Sciences should find a standard of how Bayesian analyses are to be conducted. It is of utmost importance that, at the very least, authors are completely open and transparent about what background information has been used to construct the priors. This way, we can learn from each other and find inspiration from each other and develop best practices in the field.

Third, the mathematical specification of prior distributions can be difficult. Depaoli and van de Schoot (2017) recommend a closer collaboration between substantive researchers and mathematicians/statisticians. Moreover, contemporary statistical models include many parameters. Background knowledge about some of these parameters, for example co-variances in structural equation models or class membership in latent class models, is seldom available. Thus, it is difficult to specify appropriate informative priors. In this case, non-informative prior distributions that are specified as recommended in the methodological literature are a convenient way to avoid the inclusion of unwanted information in a model. Background knowledge about substantive parameters, however, such as regression coefficients and factor loadings, should be incorporated into models whenever possible and informative prior distributions should be specified and used.

Changing how we conduct research

It is clear that the transition of educational science into a truly cumulative scientific discipline requires more than just a change in the dominant method of data analysis. Bayesian statistics is no silver bullet which magically solves each and every methodological problem. The

greatest potential of meta-analytic thinking in general and of Bayesian statistics in particular lies in changing the culture of discussion and research practice in educational science. The non-debateable black box of absolute decision-making may give way to transparent collaborative research, in which all underlying assumptions are exposed to scrutiny and discussion. An example of this collaborative research is a so-called prospective meta-analysis (Berlin and Ghersi 2005). Researchers collaborate in a programmatic series of studies in which it is agreed beforehand that these studies will serve as elements of a meta-analysis concluding the research project. Bayesian statistics is particularly well suited to such collaborative, prospective meta-analytic research. Each study serves as prior information for the next, thereby incrementally improving knowledge regarding the magnitude and uncertainty of the effect under investigation. The research project is complete once a sufficiently precise estimate of the effect is reached. In Bayesian terms, a precise parameter estimate is an estimate with a reasonably narrow highest density interval. The term "reasonably narrow" and the "relevant" magnitude of an estimate are open to debate among researchers and practitioners. Knowing that the results of existing studies can be used in subsequent analyses may lead to changes in publication policies, including a shift away from the focus on statistical significance.

Limitations of the study

Prior to concluding this review, two limitations need to be considered, both related to the focus of the literature search. First, the Thomson Reuters Journal Citation Report, Social Sciences Edition 2014 (Thomson Reuters 2015) is a dynamic list of journals. Journals are added and deleted from this list regularly; as such, it is a snapshot. The top journals, however, appear there regularly. Moreover, there may be journals that are not on this list yet present studies that include Bayesian statistics. Their impact might be considerably constrained due to their low visibility. The focus of this review on journals in the Journal Citation Report ensured that journals with the highest visibility in educational research were identified and searched. Second, it was not possible to search the complete publication histories of all journals due to limitations of the search engines (not all journals provide online archives). Thus, the prevalence of Bayesian statistics in educational research might be even lower than 0.08%. The general results and conclusions of this review, however, are not substantially affected by the omitted literature because most of the recent research (particularly that published since the 1990s when Bayesian statistics started to become more prevalent) is covered.

Concluding remarks

For Bayesian statistics to establish itself in educational research, it is imperative that researchers embark on three key epistemological shifts from frequentist to Bayesian inference (Kaplan and Depaoli 2013): (1) parameters are no longer regarded as fixed but rather as random and unknown; (2) probability becomes a concept of the degree of uncertainty and belief, which is accompanied by transparent subjective elements; and (3) a meta-analytic method of inference focused on the incremental improvement of magnitudes and uncertainties of effects and on a steady accumulation of knowledge. As has been shown in this paper, but it cannot be stressed enough, Bayesian statistics alone does not guarantee a truly cumulative research. Its success still depends on carefully developed research designs, reliable measures and valid

analytic procedures (Kruschke, Aguinis, and Joo 2012). Writing proper mathematical descriptions of empirical phenomena is still the responsibility of the researcher. For Bayesian statistics to fully unfold its potential for transforming educational science into a truly cumulative scientific discipline, a fully-fledged paradigm shift is required, affecting core policies of the discipline and involving a far-reaching rethinking of the principles of research (Andrews and Baguley 2013). The groundwork has already been laid: increasingly powerful computers have made current MCMC methods viable, numerous textbooks aim at making researchers familiar with the principles and practice of Bayesian statistics (Smithson 2010) and Bayesian statistics slowly establishes itself in the educational research literature. It is now the task of researchers to refocus their practice on the steady accumulation of knowledge via meta-analytic thinking. It is hoped that this review provides valuable input to discussions of Bayesian inference in educational research, that it contributes to a better understanding of its principles and that it points out possibilities and chances for making educational sciences a cumulative scientific discipline.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Rens van de Schoot was supported by the Netherlands Organization for Scientific Research [grant number NWO-VIDI-452-14-006].

ORCID

Christoph König http://orcid.org/0000-0003-3172-7029
Rens van de Schoot http://orcid.org/0000-0001-7736-2091

References

Studies marked with an asterisk (*) were included in the review of empirical research articles (RQ 3).

*Ageyi-Baffour, Peter, Sarah Rominski, Emmanuel Nakua, Mawuli Gyakobo, and Jody R. Lori. 2013. "Factors That Influence Midwifery Students in Ghana When Deciding Where to Practice: A Discrete Choice Experiment." *BMC Medical Education* 13: 64–70. doi:10.1186/1472-6920-13-64.

Ahn, Soyeon, Allison J. Ames, and Nicholas D. Myers. 2012. "A Review of Meta-Analyses in Education: Methodological Strengths and Weaknesses." *Review of Educational Research* 82: 436–476. doi:10.3102/0034654312458162.

Albert, James H. 1992. "Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling." *Journal of Educational Statistics* 17: 251–269. doi:10.2307/1165149.

Andrews, Mark, and Thom Baguley. 2013. "Prior Approval: The Growth of Bayesian Methods in Psychology." *British Journal of Mathematical and Statistical Psychology* 66: 1–7. doi:10.1111/bmsp.12004.

Ashby, Deborah. 2006. "Bayesian Statistics in Medicine: A 25 Year Review." *Statistics in Medicine* 25: 3589–3631. doi:10.1002/sim.2672.

*Bekele, Rahel, and Maggie McPherson. 2011. "A Bayesian Performance Prediction Model for Mathematics Education: A Prototypical Approach for Effective Group Composition." *British Journal of Educational Technology* 42: 395–416. doi:10.1111/j.1467-8535.2009.01042.x.

Berger, James. 2006. "The Case for Objective Bayesian Analysis." *Bayesian Analysis* 3: 385–402. doi:10.1214/06-BA115.

Berlin, Jesse A., and Davina Ghersi. 2005. "Preventing Publication Bias: Registries and Prospective Meta-Analysis." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, edited by Hanna R. Rothstein, Alexander J. Sutton and Michael Borenstein, 35–48. Chichester, UK: Wiley. doi:10.1002/0470870168.ch3.

*Boyd, Donald, Pamela Grossman, Karen Hammerness, Hamilton Lankford, Susanna Loeb, Matthew Ronfeldt, and James Wyckoff. 2012. "Recruiting Effective Math Teachers: Evidence from New York City." *American Educational Research Journal* 49: 1008–1047. doi:10.3102/0002831211434579.

*Buckley, Jack, and Mark Schneider. 2005. "Are Charter Schools Harder to Educate? Evidence from Washington, D.C." *Educational Evaluation and Policy Analysis* 27: 365–380. doi:10.3102/01623737027004365.

*Clewley, Natalie, Sherry Y. Chen, and Xiaohui Liu. 2011. "Mining Learning Preferences in Web-Based Instruction: Holists Vs. Serialists." *Educational Technology & Society* 14: 266–277.

Cumming, Goeff. 2014. "The New Statistics: Why and How." *Psychological Science* 25: 7–29. doi:10.117/0956797613504966.

*Cummings, Kelli D., Michael L. Stoolmiller, Scott K. Baker, Hank Fien, and Edward Kame'enui. 2015. "Using School-Level Student Achievement to Engage in Formative Evaluation: Comparative School-Level Rates of Oral Reading Fluency Growth Conditioned by Initial Skill for Second Grade Students." *Reading and Writing* 28: 105–130. doi:10.1007/s11145-014-9512-5.

Darnieder, William. 2011. "Bayesian Methods for Data-Dependent Priors." Electronic Thesis or diss., Ohio State University. https://etd.ohiolink.edu/.

Depaoli, Sarah, and Rens van de Schoot. 2017. "Improving Transparency and Replication in Bayesian Statistics: The WAMBS-Checklist." *Psychological Methods* 22: 240–261. doi:10.1037/met0000065.

*Doyle, William R. 2010. "U.S. Senator's Ideal Points for Higher Education: Documenting Partisanship, 1965-2004." *The Journal of Higher Education* 81: 619–644. doi:10.1353/jhe.2010.0006.

*Doyle, William R., and Alexander V. Gorbunov. 2011. "The Growth of Community Colleges in the American States: An Application of Count Models to Institutional Growth." *Teachers College Record* 113: 1794–1826.

*Fraile, Ruben, and Francisco Bosch-Morell. 2015. "Considering Teaching History and Calculating Confidence Intervals in Student Evaluations of Teaching Quality." *Higher Education* 70: 55–72. doi:10.1007/s10734-014-9823-0.

*Galbraith, Craig S., and Gregory B. Merrill. 2012. "Faculty Research Productivity and Standardized Student Learning Outcomes in a University Teaching Environment: A Bayesian Analysis of Relationships." *Studies in Higher Education* 37: 469–480. doi:10.1080/03075079.2010.523782.

*Galbraith, Craig S., Gregory B. Merrill, and Doug M. Kline. 2012. "Are Student Evaluations of Teaching Effectiveness Valid for Measuring Student Learning Outcomes in Business Related Classes? A Neural Network and Bayesian Analyses." *Research in Higher Education* 53: 353–374. doi:10.1007/s11162-011-9229-0.

Garcia, Patricio, Silvia Schiaffino, and Analia Amandi. 2008. "An Enhanced Bayesian Model to Detect Students' Learning Styles in Web-Based Courses." *Journal of Computer Assisted Learning* 24: 305–315. doi:10.1111/j.1365-2729.2007.00262.x.

Gelman, Andrew, and Cosma Rohilla Shalizi. 2013. "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology* 66: 8–38. doi:10.1111/j.2044-8317.2011.02037.x.

Gilger, Jeffrey W., Bruce F. Pennington, and J. C. Defries. 1991. "Risk for Reading Disability as a Function of Parental History in Three Family Studies." *Reading and Writing* 3: 205–217. doi:10.1007/BF00354958.

*Gudmestad, Aarnes, Leanna House, and Kimberly L. Geeslin. 2013. "What a Bayesian Analysis Can Do for SLA: New Tools for the Sociolinguistic Study of Subject Expression in L2 Spanish." *Language Learning* 63: 371–399. doi:10.1111/lang.12006.

*Heck, Ronald H. 2000. "Examining the Impact of School Quality on School Outcomes and Improvement: A Value-Added Approach." *Educational Administration Quarterly* 36: 513–552. doi:10.1177/00131610021969092.

Henson, Robin K., Darrell M. Hull, and Cynthia S. Williams. 2010. "Methodology in Our Education Research Culture: Toward a Stronger Collective Quantitative Proficiency." *Educational Researcher* 39: 229–240. doi:10.3102/0013189X10365102.

*Hoogerheide, Lennnart, Joern H. Block, and Roy Thurik. 2012. "Family Background Variables as Instruments for Education in Income Regressions: A Bayesian Analysis." *Economics of Education Review* 31: 515–523. doi:10.1016/j.econedurev.2012.03.001.

Ibrahim, Joseph G., Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. 2015. "The Power Prior: Theory and Applications." *Statistics in Medicine* 34: 3724–3749. doi:10.1002/sim.6728.

*Jin, Hui, and Donald B. Rubin. 2009. "Public Schools versus Private Schools: Causal Inference with Partial Compliance." *Journal of Educational and Behavioral Statistics* 34: 24–45. doi:10.3102/1076998607307475.

Kadane, Joseph B. 2011. *Principles of Uncertainty*. Boca Raton: Chapman & Hall.

Kaplan, David. 2014. *Bayesian Statistics for the Social Sciences*. London: Guilford.

Kaplan, David, and Sarah Depaoli. 2013. "Bayesian Statistical Methods." In *The Oxford Handbook of Quantitative Methods Volume 1: Foundations*, edited by Todd D. Little, 407–437. New York: Oxford University Press. doi:10.1093/oxfordhb/9780199934874.013.0020.

*Karpudewan, Mageswary, Zurida Ismail, and Wolff-Michael Roth. 2012. "Promoting pro-Environmental Attitudes and Reported Behaviors of Malaysian Pre-Service Teachers Using Green Chemistry Experiments." *Environmental Education Research* 18: 375–389. doi:10.1080/13504622.2011.622841.

*Karpudewan, Mageswary, Wolff-Michael Roth, and Zurida Ismail. 2015. "The Effects of 'Green Chemistry' on Secondary School Students' Understanding and Motivation." *The Asia-Pacific Education Researcher* 24: 35–43. doi:10.1007/s40299-013-0156-z.

Kirk, Roger E. 2003. "The Importance of Effect Magnitude." In *Handbook of Research Methods in Experimental Psychology*, edited by Stephen F. Davis, 83–105. Malden, MA: Blackwell. doi:10.1002/9780470756973.ch5.

Kirkham, Jamie J., Richard D. Riley, and Paula R. Williamson. 2012. "A Multivariate Meta-Analysis Approach for Reducing the Impact of Outcome Reporting Bias in Systematic Reviews." *Statistics in Medicine* 31: 2179–2195. doi:10.1002/sim.5356.

Kruschke, John K. 2013a. "Posterior Predictive Checks Can and Should Be Bayesian: Comment on Gelman and Shalizi, 'Philosophy and the Practice of Bayesian Statistics'." *British Journal of Mathematical and Statistical Psychology* 66: 45–56. doi:10.1111/j.2044-8317.2012.02063.x.

Kruschke, John K. 2013b. "Bayesian Estimation Supersedes the T-Test." *Journal of Experimental Psychology: General* 142: 573–603. doi:10.1037/a0029146.

Kruschke, John K. 2015. *Doing Bayesian Data Analysis*. 2nd ed. London: Academic Press.

Kruschke, John K., and Torrin M. Liddell. 2016. "The Bayesian New Statistics: Two Historical Trends Converge." https://ssrn.com/abstract=2606016. doi:10.2139/ssrn.2606016.

Kruschke, John K., Herman Aguinis, and Harry Joo. 2012. "The Time Has Come: Bayesian Methods for Data Analysis in the Organizational Sciences." *Organizational Research Methods* 15: 722–752. doi:10.1177/0956797613504966.

Lambdin, Charles. 2012. "Significance Tests as Sorcery: Science is Empirical – Significance Tests Are Not." *Theory & Psychology* 22: 67–90. doi:10.1177/0959354311429854.

*Laru, Jari, Piia Näykki, and Sanna Järvelä. 2012. "Supporting Small-Group Learning Using Multiple Web 2.0 Tools: A Case Study in the Higher Education Context." *Internet and Higher Education* 15: 29–38. doi:10.1016/j.iheduc.2011.08.004.

*Li, Xingshan, and Wei Shen. 2013. "Joint Effect of Insertion of Spaces and Word Length in Saccade Target Selection in Chinese Reading." *Journal of Research in Reading* 36: 64–77. doi:10.1111/j.1467-9817.2012.01552.x.

Mackel, Matthew C., and Jonathan A. Plucker. 2014. "Facts Are More Important than Novelty. Replication in the Educational Sciences." *Educational Researcher* 43: 304–316. doi:10.3102/0013189X14545513.

Maxwell, Scott E., Ken Kelley, and Joseph R. Rausch. 2008. "Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation." *Annual Review of Psychology* 59: 537–563. doi:10.1146/annurev.psych.59.103006.093735.

*May, Henry, and Jonathan A. Supovitz. 2006. "Capturing the Cumulative Effects of School Reform: An 11-Year Study of the Impacts of America's Choice on Student Achievement." *Educational Evaluation and Policy Analysis* 28: 231–257. doi:10.3102/01623737028003231.

*McDermott, Paul A., Marley W. Watkins, Michael J. Rovine, and Samuel H. Rikoon. 2014. "Informing Context and Change in Young Children's Sociobehavioral Development – The National Adjustment Scales for Early Transition in Schooling (ASETS)." *Early Childhood Research Quarterly* 29: 255–267. doi:10.1016/j.ecresq.2014.02.004.

Merkle, Edgar C., and Yves Rosseel. 2016. "Blavaan: Bayesian Structural Equation Models via Parameter Expansion." arXiv1511.05604. https://arxiv.org/abs/1511.05604.

*Meyer, Katrina A., and Yonghong Jade Xu. 2007. "A Bayesian Analysis of the Institutional and Individual Factors Influencing Faculty Technology Use." *Internet and Higher Education* 10: 184–195. doi:10.1016/j.iheduc.2007.06.001.

Muthén, Linda K., and Bengt Muthén. 1998-2017. *Mplus User's Guide*. 7th ed. Los Angeles, CA: Muthén & Muthén.

*Notenboom, Annelise, and Pieter Reitsma. 2007. "Spelling Dutch Doublets: Children's Learning of a Phonological and Morphological Spelling Rule." *Scientific Studies of Reading* 11: 133–150. doi:10.1080/10888430709336556.

*Petscher, Yaacov. 2010. "A Meta-Analysis of the Relationship between Student Attitudes towards Reading and Achievement in Reading." *Journal of Research in Reading* 33: 335–355. doi:10.1111/j.1467-9817.2009.01418.x.

Plummer, Martyn. 2016. "Rjags: Bayesian Graphical Models Using MCMC. R Package Version 4-6." https://CRAN.R-project.org/package=rjags.

Press, S. James. 2003. *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. 2nd ed. Hoboken, NJ: Wiley. doi:10.1002/9780470317105.

Price, Larry R. 2012. "Small Sample Properties of Bayesian Multivariate Autoregressive Time Series Models." *Structural Equation Modeling* 19: 51–64. doi:10.1080/10705511.2012.634712.

Reuters, Thomson. 2015. *Journal Citation Report, Social Sciences Edition 2014*. Rochester, NY: Thomson Reuters.

*Rowan, Brian, and Robert J. Miller. 2007. "Organizational Strategies for Promoting Instructional Change: Implementation Dynamics in Schools Working with Comprehensive School Reform Providers." *American Educational Research Journal* 44: 252–297. doi:10.3102/0002831207302498.

Savage, Leonard. 1962. *The Foundations of Statistical Inference*. New York: Wiley.

*Scarpino, Shelley E., Frank R. Lawrence, Megan D. Davison, and Carol S. Hammer. 2011. "Predicting Bilingual Spanish-English Children's Phonological Awareness Abilities from Their Preschool English and Spanish Oral Language." *Journal of Research in Reading* 34: 77–93. doi:10.1111/j.1467-9817.2010.01488.x.

Schmidt, Frank L. 1996. "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers." *Psychological Methods* 1: 115–129. doi:10.1037/1082-989X.1.2.115.

Segawa, Eisuke. 2005. "A Growth Model for Multilevel Ordinal Data." *Journal of Educational and Behavioral Statistics* 30: 369–396. doi:10.3102/10769986030004369.

*Seltzer, Michael H., Ken A. Frank, and Anthony S. Bryk. 1994. "The Metric Matters: The Sensitivity of Conclusions about Growth in Student Achievement to Choice of Metric." *Educational Evaluation and Policy Analysis* 16: 41–49. doi:10.3102/01623737016001041.

Smithson, Michael. 2010. "A Review of Six Introductory Texts on Bayesian Methods." *Journal of Educational and Behavioral Statistics* 35: 371–374. doi:10.3102/1076998610367814.

Stan Development Team. 2017. "RStan: The R Interface to Stan. R Package Version 2.15.1." https://mc-stan.org/.

*Suchodoletz, Antje von, and Catherine Gunzenhauser. 2013. "Behavior Regulation and Early Math and Vocabulary Knowledge in German Preschool Children." *Early Education and Development* 24: 310–331. doi:10.1080/10409289.2012.693428.

Sun, Shuyan, Wei Pan, and Lihshing Leigh Wang. 2010. "A Comprehensive Review of Effect Size Reporting and Interpreting Practices in Academic Journals in Education and Psychology." *Journal of Educational Psychology* 102: 989–1004. doi:10.1037/a0019507.

Taleb, Nassim Nicholas. 2007. *Black Swan: The Impact of the Highly Improbable*. New York: Random House.

*Tammets, Kairit, Kai Pata, and Mart Laanpere. 2013. "Promoting Teachers' Learning and Knowledge Building in a Socio-Technical System." *The International Review of Research in Open and Distance Learning* 14: 251–272. doi:10.19173/irrodl.v14i3.1478.

Thompson, Bruce. 2002. "'Statistical', 'Practical', and 'Clinical': How Many Kinds of Significance Do Counselors Need to Consider?" *Journal of Counseling and Development* 80: 64–71. doi:10.1002/j.1556-6678.2002.tb00167.x.

*Tobias, Justin L. 2002. "Model Uncertainty and Race and Gender Heterogeneity in the College Entry Decision." *Economics of Education Review* 21: 211–219. doi:10.1016/S0272-7757(01)00002-4.

van de Schoot, Rens, David Kaplan, Jaap Denissen, Jens B. Asendorpf, Franz J. Neyer, and Marcel A. G. van Aken. 2014. "A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research." *Child Development* 85: 842–860. doi:10.1111/cdev.12169.

van de Schoot, Rens, Joris J. Broere, Koen H. Perryck, Marielle Zondervan-Zwijnenburg, and Nance E. van Loey. 2015. "Analyzing Small Data Sets Using Bayesian Estimation: The Case of Posttraumatic Stress Symptoms following Mechanical Ventilation in Burn Survivors." *European Journal of Psychotraumatology* 6: 25216. doi:10.3402/ejpt.v6.2521.

van de Schoot, Rens, Sonya D. Winter, Oisin Ryan, Marielle Zondervan-Zwijnenburg, and Sarah Depaoli. 2017. "A Systematic Review of Bayesian Papers in Psychology: The Last 25 Years." *Psychological Methods* 22: 217–239. doi:10.1037/met0000100.

Vanpaemel, Wolf. 2010. "Prior Sensitivity in Theory Testing: An Apologia for the Bayes Factor." *Journal of Mathematical Psychology* 54: 491–498. doi:10.1016/j.jmp.2010.07.003.

Wagenmakers, Eric-Jan, Richard D. Morey, and Michael D. Lee. 2016. "Bayesian Benefits for the Pragmatic Researcher." *Current Directions in Psychological Science* 25: 169–176. doi:10.1177/0963721416643289.

*Wheadon, Christopher. 2013. "Using Modern Test Theory to Maintain Standards in Public Qualifications in England." *Research Papers in Education* 28: 628–647. doi:10.1080/02671522.2012.706631.

*Zwick, Rebecca. 1993. "The Validity of the GMAT for the Prediction of Grades in Doctoral Study in Business and Management: An Empirical Bayes Approach." *Journal of Educational Statistics* 18: 91–107. doi:10.3102/10769986018001091.