Using Bayesian statistical modelling as a bridge between quantitative and qualitative analyses: illustrated via analysis of an online teaching tool

Using Bayesian statistical modelling as a bridge between quantitative and qualitative analyses: illustrated via analysis of an online teaching tool

Samantha Low-Choy[ab], Tasha Riley[c] and Clair Alston-Knox[ab]

[a]Office of the Dean Research Arts/Education/Law, Griffith University, Mt Gravatt campus, Australia; [b]Researcher Education & Development team (RED), Office of Research, Griffith University, Nathan campus, Australia; [c]Griffith Institute of Education Research, Griffith University, Mt Gravatt campus, Australia

ARTICLE HISTORY
Compiled October 16, 2017

ABSTRACT

Bayesian methods provide a more general approach to statistical analysis that mathematically includes Null Hypothesis Significance Testing (NHST) and classical statistical modelling as special cases. This expanded, Bayesian, approach provides several benefits, which we illustrate using a case study about decision-making by teachers. We focus on a relatively unexplored topic: the way in which a Bayesian approach provides a 'bridge' between qual/quant methods. We highlight five bridges, illustrated using the case study: (1) visualization of the conceptual framework, (2) generalization via randomization and alternatives, (3) stories for interpretation, (4) computation that is flexible, and (5) continual learning, through priors. This work illustrates these bridges using a case study on a digital tool that wove together: a behavioural study to investigate decision-making, with an inbuilt perceptual component to probe rationale for specific decisions, and an interview component. A mixed method was therefore a natural choice for integrating learnings across these data sources. This online tool inhabits the digital realm of education which enables 'virtual' assessment, and sits midway between theoretical, written pieces and the fully immersed practicums in the classroom. Adopting a similar tool for training in a virtual classroom could provide greater accessibility and privacy, and enhanced feedback through the *in situ* qual/quant analytics. Thus the digital learning sphere provides a context for raising awareness of the potential that Bayesian statistical paradigm offers researchers who wish to connect qual/quant methods.

KEYWORDS
stereotyping; mixed methods; item response theory; randomization; graphical models

1. Introduction

The practice of mixed methods tends to segregate qualitative and quantitative analyses. For example, a study about teacher decision-making may be structured to comprise both a behavioural component, to observe actions, and a perceptual component, to elicit the teacher's perceptions, thereby providing insight into their underlying rationale and intentions. Using this approach, we could segregate analyses of the quantitative component of behavioural data from the qualitative components of perceptual and interview data.

CONTACT S. Low-Choy. Email: s.low-choy@griffith.edu.au

This technique could result in a qualitatively-led analysis (e.g. Hesse-Biber, Rodriguez, and Frost, 2015), denoted QUAL $\rightarrow$ QUAN + qual, indicating that an initial qualitative analysis of the interview framed the quantitative analysis of the behavioural exercise, supplemented by insights from the perceptual exercise. Alternatively a quantitatively-led analysis could be conducted (e.g. Mark, 2015), such as QUAN + qual $\rightarrow$ QUAL, indicating an initial analysis of the behavioural exercise, supplemented by qualitative analysis of the perceptual data, both generating themes for the qualitative analysis of the interviews.

In this paper we propose to explore a more thorough 'mixing-in' of the qualitative and quantitative components of analysis. We adopt a more 'systemic' perspective of how the components interact (Maxwell, Chmiel, and Rogers, 2015), rather than the 'topological' view that simply classifies the kind of design (as described above). We aim for 'genuine integration' rather than segregation where qualitative and quantitative findings are 'mutually informative' (Bryman, 2007). However, this kind of integration or 'mixing-in' of quantitative and qualitative research methods is rarely achieved (Drabble and O'Cathain, 2015). Here we consider how a Bayesian statistical approach may facilitate this mixing-in.

Interestingly, in some mixed methods texts (e.g. Hesse-Biber et al., 2015, Table 1.1), quantitative research is often considered synonymous with a particular paradigm of statistical analysis: null hypothesis significance testing (NHST), which was introduced in the 1930s. However, due to their logical construction, NHST and the associated p-values are limited regarding the insight they can provide (Wasserstein and Lazar, 2016). A key issue is the dichotomization of research hypotheses, which can lead to 'silly nulls', where an effect size of zero is not tenable, or researchers may contrive a 'straw man' hypothesis, which is almost guaranteed to be rejected (Elliott and Brook, 2007; Waller and Johnson, 1998). In contrast to *testing*, both classical and Bayesian approaches to statistical *modelling* go beyond investigating what data would be obtained only when the null hypothesis holds true, yielding a more enlightening analysis across a spectrum of hypotheses. For example, regression allows researchers to estimate the size of an effect, rather than ANOVA's focus on whether data are consistent with no effect (Gigerenzer, Krauss, and Vitouch, 2004). A classical (Frequentist) regression finds the effect size that makes the data most likely, by maximizing the *likelihood* function. This can be problematic for small samples, when due to lack of a 'clear signal', the data may be equally likely under many effect sizes, or for large samples, when due to the curse of dimensionality the data are so specific that they are unlikely to occur for any effect size (Fidler and Cumming, 2005; Greenland, Daniel, and Pearce, 2016). Bayesian regression instead evaluates the plausibility of all effect sizes. More generally, because Bayesian inference reports plausibility of hypotheses, rather than the likelihood of data under particular hypotheses, it intuitively supports evaluation of multiple working hypotheses (Chamberlin, 1890; Elliott and Brook, 2007), of particular concern for exploratory or pioneering studies.

Thus many accounts already exist of the benefits and steps involved in a Bayesian approach, however, these have tended to *focus solely on quantitative studies*. Of interest here is the way in which *a Bayesian approach can help bridge the gaps between quantitative and qualitative analyses*, both conceptually and collaboratively. Our work was motivated by an analysis of a mixed methods study, to analyse a combination of behavioural and perceptual data.

We describe five ways that a Bayesian approach to analyzing data may bridge with qualitative analyses, as outlined in Section 2. Examining these bridges gradually reveals more detail about the case study, starting with the theoretical and conceptual framework (Section 3). The Bayesian approach enables wider flexibility in how evidence can be generalised beyond the study and the resulting design of data collection (Section 4), but also enables easier communication between the qual/quant components during interpretation of the model (Section 5, which includes a brief summary of the data and the results from

the case study). Further bridges between qual/quant are provided via the flexible Bayesian computation (Section 6) and a *modus operandi* that is inherently iterative by providing a framework for accumulation of knowledge (Section 7). This methodology provides a basis for re-examining the way in which statistical paradigms shape the research (axiologically, epistemologically and ontologically), discussed further in Section 8.

2. Bayesian Statistical Modelling: Bridges for Mixing-in with Qualitative Analysis

2.1. The Bayesian Approach and its Benefits

To achieve a change in logic, the distinguishing feature of Bayesian statistics is that it reweights the *likelihood* of the data under specific hypotheses using *prior* estimates about the relative plausibility of each hypothesis. Priors may be informative, encapsulating the current state of knowledge, and obtained from previous studies, risk assessments or elicitation from experts (Low-Choy, 2013; Meyer, 1964). Alternatively, priors may equally be non-informative, reflecting complete ignorance about plausible hypotheses. Indeed a whole *community* of priors may be used to reflect schools of thought, e.g. skeptical *vs* optimistic about the potential benefits of a new drug (Spiegelhalter, Adams, and Myles, 2004). Importantly, for qualitative researchers, this provides an explicit mechanism for acknowledging the researchers' stance at the outset (Lilford and Braunholtz, 2003), although allowing comparison to the prior state of ignorance. For this reason, a first Bayesian analysis on a new problem may often take a conservative approach of using a vaguely informative prior, which can provide results that are very similar to a Frequentist analysis. Such priors can be crucial to enabling computation, especially with small datasets or complex models (van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, and Depaoli, 2017).

Many qualitative researchers may be unaware of the ways in which this introduction of *a priori* weighting of hypotheses, afforded by a Bayesian approach, expands statistical modelling. Indeed, the Bayesian statistical framework embraces both NHST and Frequentist paradigms as special cases, comparable to non-informative priors (Gelman and Shalizi, 2013). The benefits of a Bayesian approach were recommended in education decades before computation was accessible to general researchers (e.g. Diamond, 1964; Meyer, 1964), yet Bayesian approaches are still emerging in this field (König and van de Schoot, 2017). These benefits are more well-known in other research fields—such as natural language processing (Cohen, 2016), demography (Bijak and Bryant, 2016), ecology (Ellison, 1996, 2004; Hilborn and Mangel, 1997; Low Choy, O'Leary, and Mengersen, 2009), genetics (Shoemaker, Painter, and Weir, 1999), psychology (van de Schoot et al., 2017) and social science (Jackman, 2009; McElreath, 2016; Oldehinkel, 2016). As proselytised two decades ago (Western and Jackman, 1994), the Bayesian approach is well suited to analysis of data in the social sciences, being able to: detect small effect sizes from 'fuzzy' measurement; deal with small samples; accommodate multiple sources of error and complex structure in the data; and provide a logical foundation for analysis of *exchangeable* rather than fully randomized samples, such as a census of a small population or a self-selected sample. In addition, through their flexible computation, Bayesian methods sometimes provide the only viable or feasible approach to support statistical analysis of complex models, as occurred with our case study.

A detailed comparison of the two paradigms for the mixed effects model is already well-documented. For a detailed comparison of random effects modelling under both paradigms, Stegmueller (2013) compare outcomes using an illustrative example, and Hult-

ing and Harville (1991) and Browne and Draper (2006) provide a more mathematical explanation. Bayarri and Berger (2004) detail the differences and similarities between the two paradigms, and discuss the rising popularity of computation for mixed models in a Bayesian framework. In the specific context of Item Response Theory modelling, Kieftenbeld, Natesan, and Eddy (2011) presents an application for evaluating an instrument that evaluates mathematics teaching efficacy beliefs, and Finch and French (2012) compares Frequentist with Bayesian for complex IRT models, whilst Levy (2016) provides a worked through example of a Bayesian IRT within a broad overview of the benefits of Bayesian in educational research.

2.2. Bayesian as a Bridge for Mixed Methods

Here, we focus on how the Bayesian statistical modelling paradigm facilitated 'conversation' between the quantitative aspects and qualitative aspects of the case study (whose design is summarized in Section 3 and results in Section 5). To enable this systemic integration, we identify five potential points of 'bridging' between the quantitative analysis and the other qualitative analyses. These are illustrated using our study throughout this section.

The first bridge is fundamental to a modelling approach, which could be either Bayesian or Frequentist, in constrast to a testing approach.

1. **Theoretical and conceptual framework** Teacher decisions were investigated through the lens of theories in the literature. These provide the basis for a conceptual framework that underpins both the Bayesian statistical model and themes explored through Qualitative analysis.

A Bayesian approach makes the next bridge more flexible, than would be available under a Frequentist approach.

2. **Design, and potential Generalizations** specify, respectively how the conceptual framework will be measured, in a way that ensures appropriate conclusions can be drawn. Classical (Frequentist) designs rely on randomization, which may be difficult to implement in practice, or even be unethical. Bayesian analysis relies on a weaker assumption: that the individuals being studied are effectively 'exchangeable' (beyond the information recorded, for the purposes of the study), within appropriate groups. Hence exchangeability broadens the kinds of designs that can feasibly be implemented. The design also dictates how the results from this study can be extrapolated to apply more generally (e.g. to other teachers).

The following bridge also applies more generally to a modelling approach, and could be implemented with either a Bayesian or Frequentist approach.

4. **Interpretation, using stories** about how teachers were thinking about particular hypothetical student cards, as gleaned from the interview data. The quantitative model could be both understood and tested by comparing the broad outcomes and specific predictions to stories identified through qualitative analysis.

The remaining two bridges are uniquely Bayesian.

3. **Computation** in a Bayesian setting is quite flexible, and when simulation based methods are used, is more akin to a 'plug and play' approach, where the analyst can choose which modelling options to include. In contrast computation in a Frequentist setting is often tailored to each combination of modelling options, and relies on as-

sumptions that the data are sufficiently 'large' to perform calculations by making numerical approximations.

5. **Supporting an accumulation of knowledge** , via a sequence of enquiries, rather than 'one-off' analysis. The Bayesian framework allows the outcomes of this quantitative analysis to provide a prior for future analyses, when new data accrues. This changes, ontologically, the understanding of the outcomes from a statistical analysis to be: 'What have I learnt from the data? How have my prior understandings been changed? What hypotheses are plausible? How plausible are these hypotheses?' in comparison to "Are the data surprising under the null hypothesis?" under NHST or 'Which hypothesis makes the data most likely?' in a Frequentist setting.

The fourth bridge on interpretation is enabled by, and hence presented after, the third bridge on computation.

3. Bridging via the Theoretical and Conceptual Framework

Developing a theoretical framework can help provide the basis for the conceptual as well as the mathematical model for a Bayesian analysis. Here we illustrate one way this might evolve (Figure 1).

3.1. Identifying themes of theory, via Venn diagrams

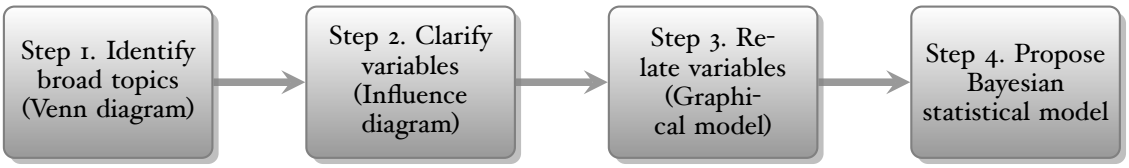

Figure 1. A pathway exploiting visualization to develop the theoretical and conceptual framework, as a foundation for statistical modelling. All steps may support any research method (quantitative or qualitative), however, Steps 2 and 3 become increasingly oriented towards a graphical model, and then more specifically a Bayesian statistical model.

Pickering and Byrne (2014) suggests that three, and sometimes four, topics are sufficient to frame most problems. Here, three underlying theories can be identified and then arranged in a Venn Diagram (Figure 2) as suggested by Pickering and Byrne (2014).

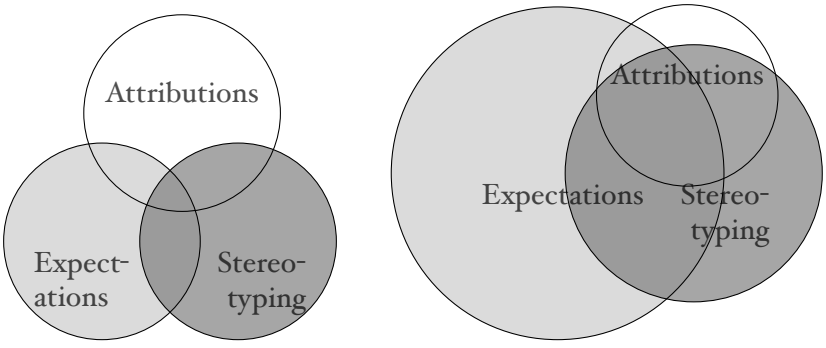

Figure 2. Theoretical framework, phase 1. Venn diagram used to represent topics for systematic literature review, following Pickering and Byrne (2014). (Left): Simple identification of three topics; (Right): Adjusted to reflect overlap between topics.

Summarizing the detailed literature review in Riley (pted), the first theory of *Stereotyping* (in dark grey) describes how people may connect characteristics or behaviours with particular groups of people, and then generalize these inappropriately (Ly and Crowshoe, 2015). The second theory concerns how teachers' *Expectations* can differ according to the student's group, which are often subject to stereotyping (in overlap between dark grey and light grey regions), e.g.: ethnicity, race, gender or other minorities (e.g. Dandy, Durkin, Barber, and Houghton, 2015). This can lead to discrimination in terms of limiting student potential and opportunities (e.g. Oakes and Guiton, 1995; Pit-ten Cate, Krolak-Schwerdt, and Glock, 2016), or specific teacher judgements (Cooper, Baturo, Warren, and Doig, 2004). *Attribution* theory (Weiner, 1974) is useful for explaining how an individual (such as a teacher) might attribute the cause of certain event (pass or failing grade of a student) to either internal (ability and effort) or external factors (task difficulty and luck) that an individual may either be deemed to control or not control. The case study focussed on the intersection of all three (the centre), where attributions may be used to explain teacher expectations and related decisions, which may be construed as stereotyping.

3.2. Clarifying concepts and variables, via an Influence diagram

Although the Venn diagram allows the researcher to organize their thoughts about collections of ideas, an influence diagram goes one step further, to clarify concepts and begin to define variables ('things' that might be measurable and hence vary). At this stage it is possible to distinguish the outcomes from the factors that influence them. Here, the teacher's decision is identified as an outcome, which is influenced by the teachers' perception of the student's previous grades, and other characteristics (bottom layer, Figure 3, left). In turn, these characteristics are influenced by stereotyping, attributions and expectations enacted by the teacher (upper layer, Figure 3, left).

Hence, the influence diagram helps distil and represent the research question:

Could we explain teachers' decisions
to place (hypothetical) students into Supplementary, Regular or Advanced learning programs

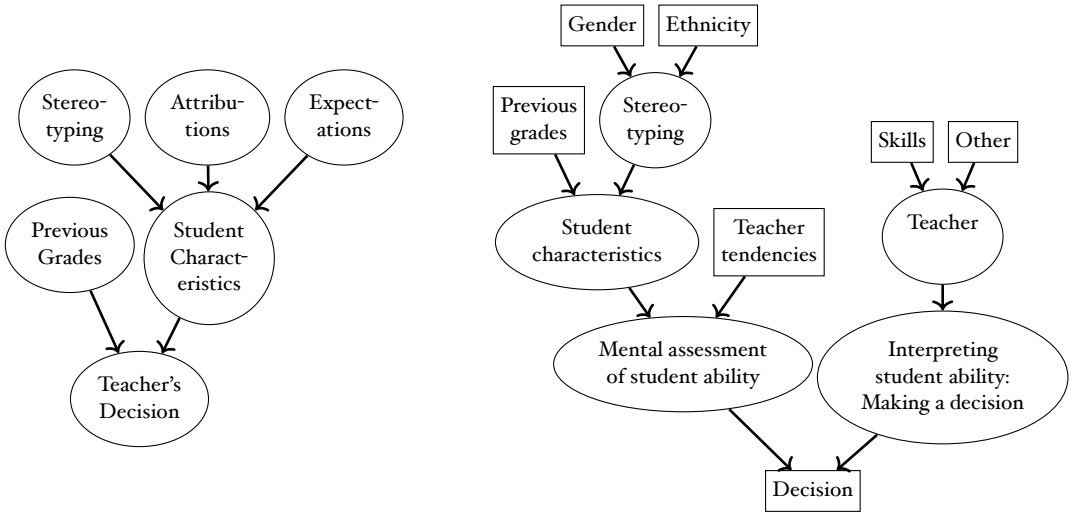

Figure 3. Theoretical framework. **Left:** Phase 1: first identify main outcome (decision) and its influences, represented using an influence diagram. **Right:** Phase 2: it is possible for a statistical modeller to expand the problem into (right) a graphical model form, which is recognizable as a Item Response Theory (IRT) model. Ellipses are concepts (not necessarily measurable), Rectangles are potentially measurable variables.

solely informed by the students' previous grades,
or were these decisions influenced by stereotyping, according to gender and ethnicity?

More specifically, the research question becomes more about quantifying effects:

What is the effect of stereotyping (e.g. by ethnicity or gender) on teacher placement decisions?

In contrast, under the NHST paradigm this stops at testing a null hypothesis of whether there are no effects

Do stereotyping effects exist, versus the alternative that they do not?

The null hypothesis would be rejected if: the set of decisions recorded in our experiment would be surprising (typically a chance of 5% or less) *when* the null hypothesis holds true. Thus NHST only considers one hypothesis in any detail: the null, and typically uninteresting hypothesis of no effect. Under the Frequentist paradigm the research question moves away from this focus on a single hypothesis. Importantly, the research question also becomes one about effect sizes rather than the existence of an effect:

How large are the stereotyping effects? Do they lead to a tendency to place students too high or too low?

The likelihood of the data is considered for every possible hypothesized effect size, and then the best estimates of the effect size make the data most likely. Whilst this considers a spectrum of hypotheses, the major limitation is the inverted logic. Many researchers mistakenly interpret the resulting p-values as the *probability of a hypothesised effect size* based on the data obtained, whereas the correct interpretation is its logical inverse:

How surprising are the data (on teachers' decisions) or *What is the likelihood of the data*, *when we fix the effect size at a hypothesized value*

This logic is not intuitive, and easily misinterpreted as its logical inverse (How likely is the hypothesis? Greenland et al., 2016), requiring careful interpretation (Wasserstein and Lazar, 2016). Under the Bayesian paradigm, the same research question about effect sizes is addressed, with the reverse logic, which is more intuitive:

How plausible are these effect sizes, for stereotyping? and hence *What are the most plausible effect sizes, for stereotyping?*.

3.3. Measurement, via a graphical model

Visualization in a graphical model more precisely defines the way in which these influences can be measured, and provides a representation of the researcher's 'mental model', which shapes both their research questions and analysis (Figure 3, right). This simplification is a crucial precursor to quantitative analysis, and also helps the researcher to refine and focus their research questions, and the underlying concepts. This 'Pre-Quantification' stage therefore requires specification of a model structure, defining variables (shapes) and relationships among them (arrows). Here concepts (circles, ellipses) are distinguished from their measurement (rectangles).

3.4. On Bayesian and Frequentist approaches

At this stage, it is not obvious that this approach is either Bayesian or Frequentist. However, during the interpretation phase, it will become evident that this visualization of the conceptual model (Figure 3) is a simplification of the graphical representation of a Bayesian model (Figure 7). In the Frequentist arena, the intuitive graphical representation

Structural Equation Models (SEM) likely explains their popularity. We note that either Frequentist or Bayesian estimation may be applied to SEM, which is actually an 'umbrella' of modelling approaches that includes IRT. Indeed Figure 3(right) could be used to define an SEM. We note that the main difference between a pathway analysis via SEM and the IRT conducted for this case study was:

> *A pathway analysis often focuses on whether the relationships (among measurements and concepts) exist, similar to ANOVA, whereas an hierarchical or mixed effects regression model (such as IRT) focuses on estimating the effect sizes involved in these relationships.*

4. Bridging, via Generalization Strategy and Design

The standard presentation of quantitative study designs is typically cursory regarding the intended use of study results. This may be partially attributable to the logical framework of NHST and p-values (Wasserstein and Lazar, 2016). However, it also reflects the predominance of one-off analyses, where only one stage of analysis is envisaged. When a researcher wishes to conduct a sequence of studies, then there are wider considerations and options regarding the ways in which generalizations can be made from each study.

There are several strategies for generalizations possible in a mixed methods study (Onwuegbuzie and Hitchcock, 2015). Each of these is closely related to the statistical design of the quantitative aspects. Here we address analytic, external and internal generalizations. We also consider epistemic uncertainty in the model itself.

4.1. Analytic generalization: from laboratory to reality

In the process of training teachers, it can be helpful to provide opportunity for assessment, which is more realistic than a theoretical exercise, yet more controlled than a full-blown classroom. For this reason the study was designed around a virtual classroom experience, where teachers were asked to make decisions about virtual students. It could be incorporated into teacher training, either within a degree program or as professional development, and could be included into a course, sandwiched between theoretical and real classroom exercises.

An online tool made it possible to create a controlled, 'virtual' classroom situation, to examine how teachers applied decision-making in the virtual classroom. At the core of any generalization is a decision of what is measured, as defined by the experimental unit (Section 4.1.1) and the experimental outcome or response (Section 4.1.2). Then the qualitative aspects focussed on eliciting the relevant perceptions, to particular decisions or overall.

4.1.1. Experimental unit

The tool introduced teachers to fictional student record cards, one at a time. The aim was to assess whether a placement decision, for each card, was consistent with the (hypothetical) student's previous grades. Of interest was whether the teacher showed any systematic tendencies related to other characteristics of the hypothetical student, as related to: stereotyping, expectations or attributions. The experimental unit was therefore a placement decision, pertaining to a specific teacher and student. With 46 teachers and 24 students, this provided 1104 units for analysis. These units can be considered *conditionally independent*, when both teacher and student are accounted for in the model. This satisfies assumptions for a Frequentist or Bayesian statistical model.

4.1.2. Experimental response

Teachers were invited to place each fictional student record card into one of three separate categories: Supplementary learning assistance, Regular, or Advanced. Hence the virtual classroom provides a context for the conceptual framework depicted in Figure 3. Qualitative researchers may be less familiar with this kind of *analytic* generalization, where a laboratory or virtual laboratory simplifies reality, in order to better control and observe the phenomenon. Although all kinds of quantitative analysis rely on this kind of generalization (Frigg and Hartmann, 2017), the model-based approaches makes key aspects of reality more explicit (e.g. through the conceptual framework), and hence suggests which aspects are to be considered less relevant.

4.1.3. Qualitative study: Probe

As noted above, teachers were asked to place students into one of the three placement categories based on their grades alone. The online tool detected and noted decisions that diverged from this guideline. At the end of the Quantitative exercise, the teacher was prompted to explain these divergent decisions, as a way of determining what factors had influenced their decision-making. For example, if a teacher placed a student of Medium-High grades into an Advanced level classroom, the teacher would be asked to explain their reasoning for this decision. Similarly, if a teacher placed students with identical grades into different programs, the teacher would be asked to explain their reasoning behind the discrepancy.

4.1.4. Qualitative study: Perceptions

Teachers were also asked to 'think aloud (van Someren, Barnard, and Sandberg, 1994) during their decision-making process. Both their movements on the screen and their verbal reflections were recorded via Camtasia software (TechSmith, 2016) for later analysis, and verification. Interviews were recorded, transcribed, and shared with teachers for verification and approval. Approved transcripts were uploaded into NVivo software (e.g. Bazeley and Jackson, 2013) where labels and codes were created and applied using an inductive and emergent process (Creswell, 2012; Merriem and Tisdell, 2015). Overarching themes were identified using Attribution theory and Wiener's three dimensions of causality (locus of causality, stability and controllability) whereas sub-themes emerged directly from the data (as detailed in Riley, pted). Sections of interviews identified with the same label or codes were cross-examined to ensure the code was relevant.

4.2. Internal statistical generalizations, within the dataset

4.2.1. Manipulated factors

Students were assigned grades at four distinct levels: Low, Medium-Low, Medium-High, and High. Two key student characteristics were manipulated. Two genders were considered: male and female. Three ethnicities considered potentially subject to stereotyping in the Australian context were: Australian Aboriginal or Torres Strait Islander, English as a Second Language (ESL), and other (predominantly Anglo-Saxon). Altogether 24 cards were presented to each teacher, to ensure representation of all 4 grades $\times$ 2 genders $\times$ 3 ethnicities.

4.2.2. Randomization of factors

In this study, it was important that the hypothetical student cards were encountered by teachers in a way that would reduce 'carryover' effects between consecutive decisions. For instance, if students were grouped according to grade, and within grade by ethnicity, then this would encourage teachers to compare the two genders when all other factors were held constant. Alternatively if students were grouped by gender then grade, then this would encourage teachers to compare ethnic groups. Thus randomization of the order in which teachers encountered hypothetical students was implemented to interrupt such patterns. See further discussion on randomization below (in Sections 4.3.3–4.3.6).

4.2.3. Controlled factors

On each card, these characteristics were communicated using a subtle clue about the students' background, ethnicity and gender: implied by name and the kind of funding previously allocated to the student. To ensure that the assessment exercise was appropriately focussed, these student cards were kept to this minimal amount of information, in order to eliminate a raft of extraneous factors that might be encountered in a 'real' world context, but distract from the purpose of the exercise, such as: the teacher's broader evaluation of a student's abilities, relative to opportunity, their home situations, and so on. This kind of control helped reduce the amount of extraneous variation that had to be explained by a model. It exemplifies the difference between a qualitative study, that considers the detailed specifics of a few individuals but may be hard to generalize across many individuals, and a quantitative study, that perceives each individual in a less specific way that therefore can be more easily generalized across many individuals.

4.3. External statistical generalizations, beyond the dataset

Generalization of the results on a statistical basis can be achieved via randomization. However, what is not always acknowledged, is that randomization is desirable because it satisfies assumptions of particular statistical models, and hence simplifies the underlying mathematics. A key assumption of many statistical methods is that the data comprise a set of independently and identically distributed (*i.i.d.*) samples: individuals are *independently* sampled from the same population, and an *identical* model can be applied to all sampled individuals. However, although randomization provides many benefits, what is little known or understood is that: *Randomization is not essential for ensuring that the analysis is useful, nor is it essential for ensuring that statistical inference is possible*. Understanding the ways in which randomization have or have not been achieved, often leads to better qualification of the interpretation of the results. This is reflected by the famous adage: 'All models are wrong, but some are more useful than others.' (George E. P. Box).

Here we address randomization and the related concept of exchangeability in general, focussing on how these ideas apply here. Accessible descriptions of exchangeability are hard to find, although practical descriptions, related to earthquake data and health care data, is provided by Draper, Hodges, Mallows, and Pregibon (1993, Sections 5.2, 5.3). In particular, Mislevy (1994) is one of few texts returned by a Google search on 'education' and 'exchangeability', and says:

> *Conditional independence also plays a key role in justifying the use of mathematical probability-based reasoning for real-world problems. The layman unfamiliar with probability and statistics, other than through informal notions about random sampling and large samples, might question whether mathematical probability has anything to do with real-world observations that are governed by disparate mecha-*

nisms and may be linked with one another in unknown ways ... Even if we admit the possibility, indeed the inevitability, of such differences among the antecedents of potential observations, yet at a given point in time have no information to distinguish among them a priori, then these observations are 'exchangeable' from our point of view. That is, our subjective probability distribution for their scores would be the same under any permutation of the variables.

4.3.1. Replication, across participants

For this study, 46 teachers were enlisted through invitation sent to the principals of primary and secondary school in three regions of Queensland, Australia. This approach is advisable and respectful when working in this educational context, and also provides a credible avenue for enlistment. Participation was self-nominated; and it would not have been advisable to impose participation on randomly selected participants, as this would introduce other biases. In contrast, it would be feasible to collect a census of a small, well-defined population (Western and Jackman, 1994), for example, setting this exercise as an assessment item for all teachers enlisted in a course on social justice in teaching. Thus, strictly speaking, this particular study of 46 teachers follows a quasi-experimental design, due to the difficulty of imposing randomization on selection of participants.

4.3.2. Exchangeability

By definition[1], exchangeability can be practically interpreted to mean that the model for all individuals does not depend on any information about which individuals were selected (e.g. the order in which they 'arrived', or groups of individuals). Hence, crucially, this self-nominated sampling process can be considered acceptable here since there was no desire to generalize study findings to all teachers in Queensland. Instead, relevance of findings can be expressed in terms of a more specific setting, of interested teachers in Queensland (Kelle, 2015). Within this set of interested teachers, those enlisted may be considered *exchangeable* with all other teachers in Queensland, who might be sufficiently motivated to enlist, and hence are interested in improving their knowledge about teaching practice in terms of appropriate decisions in relation to student assessment, and how to apply this in their teaching. Exchangeability is sufficient to support Bayesian analysis, whereas Frequentist analysis demands a stronger assumption of randomization (which satisfies exchangeability).

In many guidelines on statistical methodology, completely randomized selection of participants is presented as a well-established sampling strategy that reduces inadvertent biases such as selection, recruitment and participation biases (e.g. McCambridge, Kypri, and Elbourne, 2014; Moynihan, Lewis, Hall, Jones, Birtle, and Huddart, 2012). However it is also well-acknowledged that in many situations, where recruitment requires willing participants, it can be very difficult to deploy randomization: in social science, Western and Jackman (1994), and in risk assessment Draper et al. (1993).

4.3.3. Randomization and Bias

Here, the tool was programmed so that each teacher experienced perceptual questions that were linked to their answers for the behavioural exercise. Because of constraints of software at the time, the tool had to apply the same order of student cards for all teachers. Thus the ordering of cards was randomized, but just once, being the same for all teachers.

[1]Lindley and Novick (1981) define exchangeability mathematically as: 'A set of units is exchangeable in X, given Y , if $p(X_i = x; \text{all} i | Y_i = y_i; \text{all} i)$ is invariant under relabelling of the units.'

This double-blinding of card order (blinded to teachers, and blinded to the experimenter) helped ensure objectivity, since the experimenter did not 'choose' the ordering.

4.3.4. Randomization and Independent responses

For online tools, randomization also enables individualization of question order, hence reducing the chance of collusion and ensuring responses are independent. Here, there was low expectation of cheating or sharing answers among teachers, who were geographically dispersed; particularly since teacher access was restricted via online access to a secure website. In addition, the randomization of card order reduced the potential for a 'learning effect', which would occur if cards were ordered from 1 to 24, by grade, gender and ethnicity. With such a clear ordering, it is likely that teachers could easily detect that several student cards had the same grades, but varying gender and implied ethnicity.

4.3.5. How Inference accounts for Randomized vs Fixed orders

Because the same card ordering was used, Bayesian analysis of the group of teacher responses is predicated on this particular (fixed) order of student cards. In fact, by fixing this order, we obtain a consistent basis for comparing each teacher's decisions. However, this form of randomization does leaves open the possibility of different findings resulting from different orderings of cards.

A completely randomized design would have allocated the card-orderings randomly to each teacher. There are in fact $24! = 620 \times 10^{21}$ or 620 sextillion possible orderings, and hence there is no practical chance that all of these could be assigned in an experiment. Thus randomizing the choice of orderings ensures an unbiassed choice of which of the 620 sextillion possible orderings are chosen. In a Frequentist analysis, the data would be considered a 'pseudo-experiment' comprising 'observational data', and in strict violation of the requirement of *iid* samples.

4.3.6. Randomization and the Logic of statistical inference

If the ordering of student cards had been randomly assigned to teachers then it would be possible to appeal to Frequentist sampling properties, to support a classical analysis. The corresponding interpretation of confidence intervals (and the related concept of p-values) relies on randomization (Wasserstein and Lazar, 2016): In the long run (for a large number of random samples, of decisions), a 95% confidence interval for an effect (such as the effect of a student's grade, gender or ethnicity on a teacher's placement decision) 'hits' (or covers) the true effect size for 95% of random samples. Our random sample is just one possible, and hence it either contains the true effect size or it does not, with a 95%-vs-5% hit-or-miss rate, being the chance of hitting or missing the true effect size.

Most classical statistical analyses (in standard statistical packages) use a numerical technique called 'Maximum Likelihood Estimation' (MLE) which methodically constructs a thought experiment around a description of uncertainty called the 'likelihood function'. MLE considers: *How likely are the data, and hence how surprising are the data, for each possible value of a particular effect size? Consider all these possibilities and choose the one that makes the data most likely or least surprising.* This seems sensible, but encounters difficulties when: (a) the data are surprising under most effect sizes, which often occurs for very large datasets, where the effect size can be estimated very accurately; or (b) the data are not surprising under almost any effect size, which often occurs for small or 'noisy' datasets. This logical conundrum has been recognized since the invention of null hypothesis testing (e.g. see discussions in Fidler and Cumming, 2005; Greenland et al., 2016).

In a Bayesian context, randomization satisfies the assumption that individuals (here decisions for a teacher-student combination) are essentially exchangeable, after we account for the teacher and student. Uncertainty is typically communicated via 95% credibility interval for each effect (or combination of effects), which has more direct interpretations: it contains the 95% most plausible values.

4.4. Epistemic modelling uncertainty

All modelling exercises are inherently limited to a perspective that is framed by the model that is used. To overcome this limitation, we may broaden this perspective, and either fit different kinds of model (within a quantitative paradigm) or consider mixing-in qualitative components, as considered here.

Incorporating the perceptual component aimed to bring multiple benefits. Pedagogically, it provides an opportunity for the teacher to reflect and reconsider their decisions (although they were not able to revise them). It improves the quality of data collection, by applying a subtle form of 'quality control'. Most importantly, it elicits a 'story' (considered below) that may help explain decisions, and help 'ground-truth' or 'triangulate' outcomes from modelling.

In addition, an **interview component** followed the online study. The researcher invited teachers to respond to a series of questions in relation to their own decision-making as well as in relation to their perceptions of teachers' expectations towards certain groups of students. This interview included questions such as: 'In addition to grades, what additional factors might influence your decision to place a student into a [supplementary] learning assistance class/advanced classroom' as well as 'Do you think teachers' have different expectations of students in particular groups, [Aboriginal Torres Strait Islander / English-as-a-Second-Language / neither or Male / Female]? Why? Why not?' Teachers' responses to these questions provided further insight into the reasoning behind teachers decisions and recommendations about the educational opportunities afforded to students and are discussed elsewhere (Riley, pted).

4.5. On Bayesian and Frequentist approaches

The first two strategies of generalization form the foundation of designing data collection for any statistical analysis(see general principles of design, e.g. in Box, Hunter, and Hunter, 2005). The first strategy of analytic generalization, from laboratory to reality (Section 4.2), is generally covered by the choice of experimental unit. The second strategy of internal statistical generalizations involve manipulation, randomization or control of factors of interest. These provide a basis for the third strategy of external statistical generalizations provides a contrast of the two paradigms: exchangeability, at the heart of Bayesian inference, versus randomization, which is considered essential for a Frequentist analysis. The fourth strategy of addressing epistemic model uncertainty is achieved via mixing-in the research methods.

5. Bridging via Interpretation: Qualitative or Quantitative

5.1. Bridging via Interpretation: Qualitative component

To illustrate how the qualitative data can add meaning to the quantitative analysis, we first provide an indication of the data (Section 5.1.1) and the results (Section 5.1.2) before

demonstrating the way that qualitative data can enrich their interpretation (Section 5.1.3).

5.1.1. Quantitative Data, Behavioural study

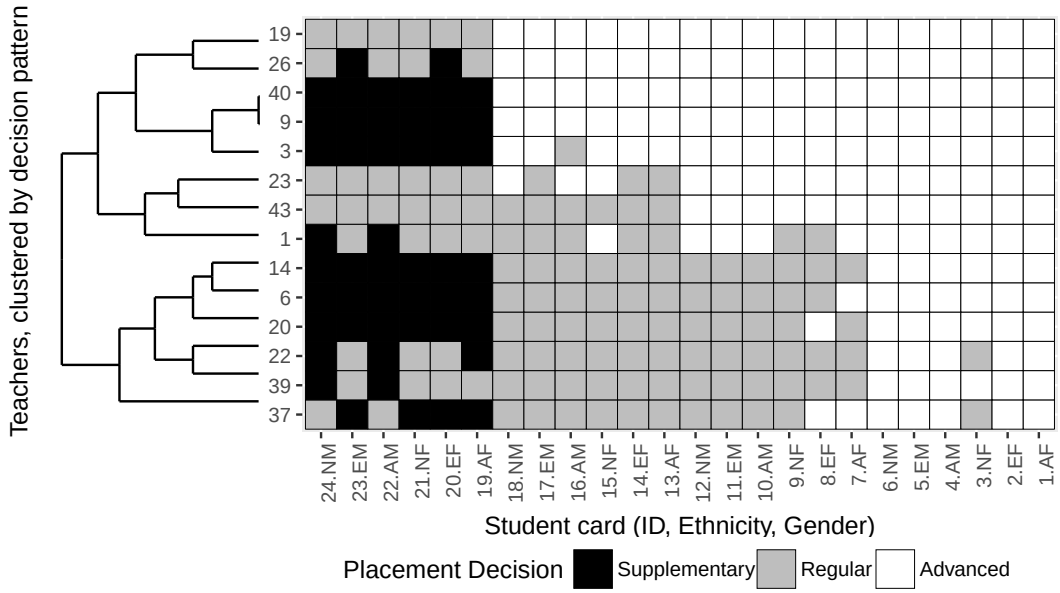

Figure 4. Data obtained from behavioural study. Student cards (columns) are presented in decreasing order of grades (L: students #19-24, ML:#13-18, MH:#7-12, H:students #1-6), ethnicity (A, E, N=non) then gender (M, F). Thus the first few students are #24-H-Non-M, #23-H-ESL-M, #22-H-Aboriginal-M, etc. Teachers (12 out of 46, shown in rows) are clustered based on their pattern of decisions (black=Supp, grey=Reg, white=Adv) across student cards (columns).

Figure 4 shows placement decisions made by a selection of teachers about a specific student: Supplementary, Regular or Advanced. Each of the fictional student records was considered by every teacher. For the purposes of illustration, we show all results for twelve teachers, selected at random from the full set of 46 teachers. We also added teachers #26 and #39 for cross-referencing with Section 5.1.3 below. The quantitative data are fully presented and explored in Riley and Low-Choy (tted).

Students are grouped by grades, then by gender and ethnicity (as designated by the colours at the top). Thus cards #1, 2 and 3 (from the right) are all female with High grades, and their ethnicity is respectively: Indigenous Australian (Aboriginal or Torres Strait Islander), English-as-a-Second-Language (ESL), or non-Indigenous and non-ESL.

We see a variety of pattern of responses across teachers, comparing rows. Teachers are clustered together if their placement decisions are similar. The dendrogram at left identifies two major groups of teachers. The lower group generally used all three categories, that was perfectly aligned with grades for Teacher #14 (top, lower group), but increasingly less so towards Teacher #37 (bottom, lower group). Teachers clustered into the upper group tended to avoid a category, either the Regular category (#3,9,40, middle, upper group) or the Supplementary category. In the latter group, some teachers reserved Regular for all Low grade students, and assigned everyone else to Advanced (#19,26, top, upper group). Alternatively when avoiding the Supplementary category, students with Low and Medium-Low grades were assigned to Supplementary and the remainder to Advanced (#23,43, bottom, upper group). The most similar teachers are more closely linked by the dendrogram: Teachers #9 and #40 allocated all L students into Supplementary and all ML, MH, H students into Advanced.

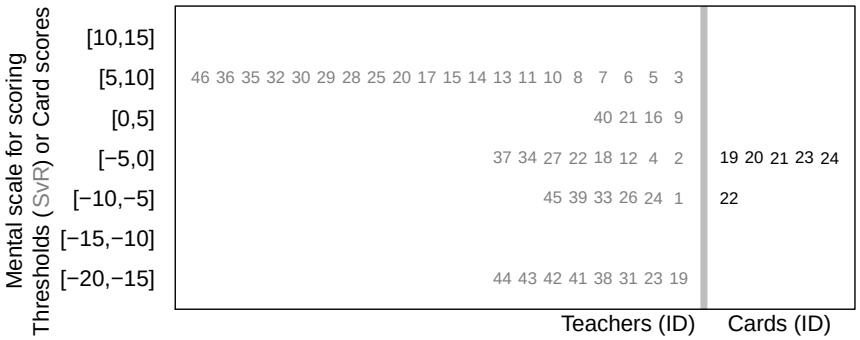

Figure 5. The mental scale inferred by the model (grey horizontal line, and x-axis) goes from -20 points (at left of x-axis) to $+60$ points (at right of x-axis). To the right of this line: we show scores given by teachers, on average (black), to each student card (black number). To the left of the line: we show thresholds used by teachers, across students, to separate Supplementary from Regular placements (grey) for each teacher (ID, grey text). Teacher IDs are stacked like a histogram, counting up the popularity of each threshold.

Although not shown here there were six teachers who gave placement decisions in alignment with grades.

5.1.2. Analysis of Quantitative Study

A detailed explication of each of the effects sizes obtained through a Bayesian analysis is provided in Riley and Low-Choy (tted). [2]. Here we focus on general patterns in the outputs[3]: the model's predictions, as summarized across teachers for particular cards, or across students for particular teachers (Figure 5). These rely on estimates of the way in which teachers, overall, place each student (above the horizontal grey line), as well as the thresholds that are used by individual teachers (below the horizontal grey line) to separate students into each category. Here we show only the thresholds between Supplementary and Regular. See Riley and Low-Choy (tted) for all thresholds, including Regular and Advanced.

In Figure 5 we show the lower half of teachers' mental scores that fall below 5 points, from a range of -10 points (for card #22) up to 60 points (not shown; see Riley and Low-Choy, tted). We also show the model's predictions of each teacher's cutoff, that separates scores for Supplementary and Regular placements. On average, we notice some trends across these placement trends. One card 'stands out': #22 scored lowest among Low grade students—he was male and indigenous. The mental score of 5 points was used by 20 teachers (e.g. #3, 46) to separate Supplementary from Regular. Several teachers (e.g. #19, 23, 31, 38, 41–44) allocated Low grade students to the Regular program, since they had a very low cutoff between Supplementary and Regular.

Interestingly, there was a similar amount of variability among teacher's idiosynchratic

[2] Essentially, previous work fit a generalized linear mixed effects model (GLMM) with a cumulative logit link for the ordinal response (Supplementary $<$ Regular $<$ Advanced), with a latent variable for the teachers' mental scores. The expected value contrasted: student cards, adjusted for grade, stereotyping effects within grade and a card-specific random effect; versus thresholds, as fixed effects specific to teacher. A 'guessing' parameter for each teacher was included as a fixed effect to reflect how consistently they applied their own thresholds. Vaguely informative priors enabled Bayesian computation, and reflected a prior state of ignorance that is similar to a classical stance. Priors for all fixed effects were Normal with large standard deviations (10). Priors for the variance components were truncated Normals on the scale of the standard deviations, as per Gelman (2006).

[3] Computation was implemented using the WinBUGS (Spiegelhalter, Thomas, Best, and Lunn, 2003) language via R (R Core Team, 2017). Convergence of the MCMC chains to the desired posterior distributions were checked using the `boa` library (Smith, 2007) in R visually using traces and the Gelman-Brooks-Rubin statistic to assess sensitivity to starting point, the Raftery-Lewis diagnostic to assess chain length, and the Hiedelberger-Welch statistics to assess stationarity.

choice of thresholds. There was some consistency in choosing the threshold between Supplementary and Regular, with one very popular choice (at 5 points), which corresponded to the choice of teachers whose decisions were determined solely by grades (#10,5,13,14,29,46).

5.1.3. Linking Stories to Results

Although this methodical breakdown of the predictions helps illustrates the outcomes of the model, it can still leave a qualitative researcher (and some quantitative researchers) wondering whether the statistical model performs well 'on average', but never performs well for any individual decision, or particular teachers or students. In this case study, it is possible to refer back to the 'stories' obtained through qualitative analysis of interview and perceptual data to find individual stories that relate to model findings.

Here the interviewer noticed a general trend in the teacher's placement decisions.

> Interviewer: Okay, so this is Paul in a regular class. And you've placed [Min] Lee in the supplementary learning. And I was just wondering why you've made that difference?
>
> Teacher (#26): For a number of reasons, [...] But more it was the link to the ESL. I was worried it wasn't, whether it was an understanding for explicit, very borderline for [Min] Lee. I really wanted to put her into the general program but, with her ESL link, I was thinking, well, grade eight, put into the program where things are very, very explicit, review for grade nine and she may be able to go back into the standard program with confidence. For this one, [...], being a boy, 'cause they're both boys, but being a boy who's not ESL, I was hoping that staying with his peers would encourage him to stay confident and keep going with that. So basically it was the ESL thing for me was a concern. If he had been ID'd as ESL as well, if Paul Dennis Kelly had been ID'd as ESL as well, I would have immediately put him into the other program. That's my only argument for it.

This confirms that ethnicity as well as gender did affect this teacher's decisions, and when prompted they were able to explain that reasoning.

In the data (Figure 4), Teacher #26 clearly shows a different pattern for allocating ESL students, in line with grades, into Supplementary yet promoting other students (including Aboriginal students) to the Regular program. This teacher did not differentiate between Medium Low, Medium High and High grades, assigning all students (regardless of ethnicity) to the Advanced program. This teacher was one of six whose thresholds hovered around student #22, a male Aboriginal (Figure 5).

Another teacher was unaware that they had placed two boys in Supplementary, and girls with the same grades into Regular. With one boy in particular (Micky), their reasoning involved both gender and ethnicity:

> Interviewer: Now you placed Paul in Supplementary, and you placed Karen in Regular, so I was just wondering why you placed these two identical cards differently?
>
> Teacher (#39): Oh no, the same! And they are the same. Look at that. Oh, I did too. Maybe that's a gender bias, possibly, I don't know.
>
> Interviewer: Okay, and yes, the other one that you had placed in Supplementary learning was Micky.
>
> Teacher: And the same? Yeah, exactly the same. Maybe I do have a gender bias when it comes to Year 7's into Year 8. I don't know. [...] I think, did I speak about him being ATSI and that transition into high school can sometimes be more difficult and that being a D, and often that can be because of ESL, especially in North Queensland, that they speak the

Creole at home so there's an aspect of ESL in their English?

Referring to the data (Figure 4, row 39), for students with Low grades, the teacher clearly assigns males (non ESL) to Supplementary category. However, all female students and the male ESL student were promoted to the Regular category. This teacher assigned all other students in line with their grades: Medium Low and Medium High students were placed in Regular, and High grade students were placed in Advanced. Thus their stereotyping is only operating at the lower end of the spectrum. Similar to teacher #26, this teacher was one of six whose thresholds hovered around student #22, a male Aboriginal (Figure 5).

The stories from both of these teachers justify the modelling decision to consider the effects of stereotyping, *as operating differently for each grade*. Although not shown here (see Riley and Low-Choy, tted), the stereotyping effects were more marked for the Low grade students, with Male gender and Aboriginal ethnicity both associated with strong negative effects on teachers' mental scores (underlying their placement decisions).

5.1.4. On Bayesian and Frequentist Interpretations

The rich information from the stories attained through qualitative method provided information on the nature of stereotyping effects, both their existence, and more importantly their direction, placing students in various groups lower or higher than other students with similar grades. Thus the stories corroborate results from a model-based approach (Frequentist or Bayesian) that focus on *effect sizes*, whereas these stories add information to NHST approaches, and may also corroborate their simpler hypotheses that *stereotyping effects exist*.

If there had been no teachers with decisions solely guided by grades, and equivalently no student cards who were always allocated in the same way, then it is possible that a Frequentist model could have been fit to this data. In that situation the same insights could be gained by matching the qualitative insights to help trace how the observed data, here for particular teachers, contributed to the model predictions. In this case the predictions provided additional information, by estimating similar thresholds for discriminating Supplementary and Regular placements, for two teachers whose decisions of Low grade students depended on gender and ethnicity.

This case study illustrates the flexibility of Bayesian computation permitting such a study to be analysed. If constrained to Frequentist techniques only, then researchers may need to consider inefficient measures, (as highlighted by Zorn, 2005): *'Separation raises a particularly difficult set of issues, often forcing researchers to choose between omitting clearly important covariates and undertaking post–hoc data or estimation corrections.'* Alternatively, they may consider it necessary to redesign the study with a continuous (e.g. the mental score) rather than ternary outcome (three possible decisions). Imposing a continuous scale can place more load on the participant, and may: be more difficult to interpret, have weaker link to decision-making in a real situation, rely on more advanced numerical skills, and hence raise a different set of issues.

We note that in some cases, a Bayesian approach to estimation of Item Response Theory models (similar to but not exactly the same as the one considered here) have become commonplace, with recent usage in an educational context by Kieftenbeld et al. (2011), who matter-of-factly, without justification, apply a Bayesian rather than a classical Frequentist approach.

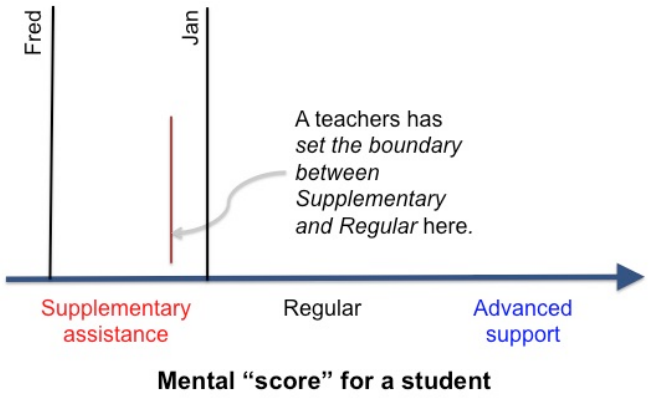

Figure 6. What the thresholds mean in an Item Response Theory model. The horizontal axis shows a continuum for any teacher's mental score. This is anchored at three distinct levels of assistance that (the teacher considers) are appropriate. Fred's score falls below the teacher's threshold between Supplementary and Regular, so he would be allocated to the Supplementary category. Jan's score falls above the threshold, so she would be assigned to Regular.

5.2. Bridging via Interpretation: Quantitative component

The conceptual framework mapped out by the graphical model can be reflected in a Bayesian statistical model. Typically this is written out as a mathematical equation. Social science researchers who are new to quantitative analysis may be more comfortable with the same equation expressed in pictures and/or words instead of algebra. Indeed, a full graphical representation is possible for any Bayesian statistical model, using a Directed Acyclic Graph (DAG), for instance using DoodleBUGS (Lunn, Thomas, Best, and Spiegelhalter, 2000). However, such graphical representations are typically explained from the mathematical standpoint (e.g. Campitelli and Macbeth, 2014), and hence not accessible to those researchers who could benefit the most. For the purposes of this paper, we focus on one aspect of the model, and demonstrate how the DAG evolves naturally from the conceptual framework (Figure 3,right), and how it relates to a 'word equation' view of the model (the last phase of Figure 1).

5.2.1. What the model is

In the previous section, we provided a conceptual framework that logically maps, how each of the factors are related to the decision outcome (Figure 3). The next step is to precisely quantify these relationships. In fact, there are many ways that this can be achieved mathematically. From the experience of the applied statistician in the research team, it was possible to recognize that an IRT (Item Response Theory) modelling framework was one suitable option. IRT encompasses the Rasch models developed to support educational testing (Rasch, 1960). It effectively separates the decision-making process into two elements:

- assessment: of each (hypothetical) student's track record
- decision-making: allocating each student to a category

The IRT model creates a way of explaining how and why the teachers' decisions are made about each hypothetical student. Imagine that teachers mentally construct a score for each student along some continuum. Here the students definitely needing supplementary or advanced assistance fall at either end of this spectrum, as depicted by the horizontal axis in Figure 6. Every teacher places each student somewhere along this spectrum. The teacher's mental score for Jan places her closer than Fred to Regular rather than

Supplementary assistance. The way that the teacher's mental scores become manifested as a tangible and measurable outcome (the decision), is enacted by the thresholds they set. The teacher's threshold distinguishing Supplementary from Regular assistance, falls between Fred and Jan, indicating that the model predicts that this teacher would allocate all students to the left of the threshold (like Fred) into Supplementary and all students falling to the right (like Jan) into Regular.

Equivalently, an IRT model can be considered as a kind of mixed effects model, also known as a hierarchical linear model (HLM, e.g. Raudenbush, Bryk, Cheong, Congdon, and du Toit, 2004), variance components model or a generalized linear mixed model (GLMM, e.g. De Boeck, Bakker, Zwitser, Nivard, Hofman, Tuerlinckx, Partchev, et al., 2011). In this instance, we find the concepts and terminology of IRT to provide a useful framework for analysis, although a GLMM would be just as appropriate. Following De Boeck et al. (2011)'s taxonomy of IRT models, all covariates have a mode, are external/internal, fixed/random, and apply to a certain scale (mode of randomness), e.g. across persons or items, potentially with nesting and/or grouping. The stereotyping covariates (grade, gender and ethnicity nested within grade) refer to students (an IRT item), with randomness across teachers. Location refers to the teacher (an IRT person), with randomness arising across students. These covariates may all be considered external and fixed effects. For each teacher, we estimate two thresholds, which are internal and fixed effects. For each teacher, we also allow a guessing parameter, which is applied with the link function.

Mathematically, a Bayesian form of an IRT model is a kind of 'graphical model' and can be represented using a Directed Acyclic Graph (Figure 7). It is a *graph* because it comprises of nodes, corresponding to variables or constructs that conceptually combine variables, and *directed* edges (i.e. arrows) that join nodes, indicating dependence of a *child* node on its *parent* nodes, and vice versa, the influence of the parents on their children. Each child node may have one or more parent nodes, and similarly each parent may have one or more children. The graph is *acyclic*, because no cycles are permitted which would violate the usual parent-child relationships, e.g. a grandchild cannot be a parent to their grandparent.

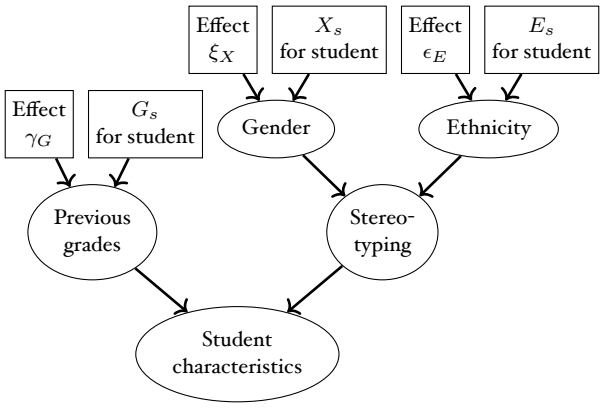

Figure 7. Expanding the graphical model of Figure 3 (right), by adding model parameters, to provide a Directed Acyclic Graph for a Bayesian hierarchical statistical model. Only one portion of the whole model is shown. A teacher's mental score of a student depends on their grade, which may be adjusted by stereotyping depending on their characteristics.

5.2.2. *How the model is used*

The IRT model underlying the results shown in Section 5 involves multiple components. A major component, which we examine in detail here, is the teacher's mental score, centred on typical scores given to each student card and hence stereotyping effects. Other components are detailed elsewhere (Riley and Low-Choy, tted): the teachers' use of thresholds between decision categories (in the graded response model portion of the model); and the teacher's tendency to consistently follow their own set of 'rules' or 'heuristics' (a guessing parameter). For the purposes of demonstrating the Bayesian bridge, we focus on one element of the model: building each teacher's *mental score* for each hypothetical *student*, *adjusting* to the stereotyping effects for each grade, but allowing each card to follow or *deviate* from the trend.

$$\begin{aligned} \text{Mental score} &= \text{Student score} + \text{Card deviation} \\ \text{Student score} &= \text{Grade} + \text{Gender adj.} + \text{Ethnicity adj.} \end{aligned} \tag{1}$$

Across teachers, student scores start at a value corresponding to the student's Grade, and then are adjusted up or down according to their Gender and Ethnicity. Six teachers did allocate students consistently with grades (e.g. #14 in Figure 4), and another five deviated on just one decision (e.g. #6 in Figure 4). The model included an 'unexplained portion' that quantified how each student card, on average, 'deviated' from an allocation based purely on grades, and accounted for some students' tendency to be scored higher or lower, compared to their grades, even after adjusting for gender or ethnicity.

This kind of model is quite flexible. In our study, this model provided scores on a 100 point scale between -20 and 80, with 0 sitting at the boundary between Supplementary and Regular category. In our study, all students allocated to the Advanced category, on average were scored above 40 on this scale, whereas teachers were much more variable in how they scored students in the lower categories (Figure 4).

Let us consider one hypothetical student card in detail. Mickey Wonaemir (Card #22) was Male and Aboriginal, with Low grades. He was one of the students who was often allocated differently to his non-Aboriginal or Female peers with the same Low grades, as evidenced by teachers #26 and #39, described in Section 5.1.3. Across teachers, on average, he was assigned the lowest score, between -5 and -10 points, in comparison to other Low grade students, scoring between 0 and -5 points (Figure 5).

We also find that Mickey was the only student who was allocated higher than expected if we consider stereotyping by gender and by ethnicity as operating independently:

$$\begin{aligned} \text{Student score} = {} & \text{Grade} + \text{Gender adjustment} + \text{Ethnicity adjustment} \\ & + \text{Combined Gender-Ethnicity adjustment} \end{aligned} \tag{2}$$

In his case we see that his male gender and Aboriginality were magnifying the stereotyping effect. In future students like this could be accommodated by adding an adjustment for the combined effect (interaction) of gender and ethnicity. However to confirm the need for this adjustment for an interaction (or moderating effect), more extensive data would be required, with more replication of hypothetical students with the same attributes, and/or across teachers, to confirm that this is not an anomaly, perhaps related to the cards seen immediately before and after.

5.2.3. *Under the hood: Effect sizes, uncertainty and interpretation across paradigms*

Because this is a statistical model, these adjustments (effects) are estimated with uncertainty. This enables us to evaluate specific research hypotheses. However, these evaluations are logically quite different across the paradigms of Bayesian, Frequentist and NHST statistics.

In a Bayesian setting, we can evaluate the plausibility of any statement about the elements of the model (the grade, and adjustments for gender and ethnicity). We continue to focus on stereotyping relevant to male students, and consider gender stereotyping in Mickey's grade. The 95% credible interval for the male adjustment to the score (compared to a baseline of females who are non-ESL and non-Aboriginal) spanned from $[-5, -1]$ for L students, compared to $[-3, +1]$ for MH students. This means that teachers plausibly scored female students higher than their male counterparts in both the L and the MH grades.

It wasn't possible to fit this model within a Frequentist setting, because of 'perfect predictability': some teachers assigned all students purely based on grades, and some other teachers avoided allocating any students to the Regular or the Supplementary categories. This is also known as the 'separation' effect, where Frequentist computation (e.g. via maximum likelihood) may fail or 'provides implausible estimates' (Rainey, 2016). However, if the resulting 95% confidence interval had been calculated to fall in the same range of $[-3, +1]$, then its interpretation is different, in fact the logical reverse (Sedimeier and Gigerenzer, 2001), and much less straightforward (Wasserstein and Lazar, 2016), and often confusing (Greenland et al., 2016). To satisfy the assumptions for Frequentist analysis, we would need to imagine many, many similar samples of teachers, randomly selected from the same population of self-nominating teachers in Queensland, volunteering for research on teacher decision-making. We would then imagine fitting the same model to each sample, and calculating the same confidence intervals. Then 95% of these confidence intervals would contain the 'true' adjustment needed for gender, which would (in the ideal situation) be calculated from a census of all such teachers. Then the correct interpretation is that there is a 1 in 20 chance that our estimated confidence interval for the adjustment for Gender does not include this 'true' value. To reiterate:

> The Frequentist confidence interval has a hit-or-miss 95% chance (compared to other random samples) of containing the (single) true value of the gender effect, whereas the Bayesian credible interval delimits the 95% most plausible values of gender effects based on the data observed, and integrates a consideration of both sampling variation and a priori estimates.

If using an NHST approach, we would note that the 95% confidence interval contains the value zero, and conclude that there is inadequate evidence to reject a 'null' hypothesis that females and male students with Medium-High categories are on average scored any differently. This contrasts with

> the Bayesian findings that most plausibly, females with these grades are allocated (slightly) higher scores than males.

5.2.4. *On Bayesian and Frequentist paradigms*

The Directed Acyclic Graph (DAG) provides a graphical representation of the Bayesian model, and is not available to Frequentist models (except for SEMs, as mentioned previously in Section 3).

The kind of IRT model used here requires simultaneous estimation of both teacher effects, including thresholds, as well as student-specific effects, related to stereotyping (Section 5.2.2). As noted in the next section, this cannot be achieved without Bayesian

computation (Section 6). The more complex model provided the depth needed to link stories with model findings (Section 5).

Finally, the interpretation of uncertainty differs logically across the Bayesian and Frequentist paradigms, as explained in Section 5.2.3.

6. Bridging via Computation

Bayesian statistical modelling can provide valid results for small sample sizes, as low as a single observation. This is enabled by specification of a prior, representing the researcher's *a priori* knowledge about the model. A classic example in medicine is where a new test for a rare genetic disease returns its first positive after 10,000 negatives: Bayesian statistics permits analysis (of rare events in binomial trials), but NHST and classical techniques fail (Suess, Gardner, and Johnson, 2002). As noted by van de Schoot et al. (2017, p226) Bayesian methods have performed well in many situations, including IRT on small samples:

> In general, Bayesian methods were found to outperform other estimation methods for many different performance criteria that is, Type I error rates, power, and producing stable and accurate coverage rates) across a wide range of statistical models (e.g., regression, CAT, IRT, SEM). These findings were especially relevant when the sample size was small (we discuss the issue of small samples in a separate subsection), when the model was more complex (e.g., Wang & Nydick, 2015), or in the situation where alternative methods were simply not developed yet (e.g., Wollack, Bolt, Cohen, & Lee, 2002).

Bayesian methods provide computational advantages in many other settings, where Frequentist or NHST methods are infeasible. Indeed Royle and Dorazio (2008) reluctantly adopt Bayesian methods specifically to enable computation of complex hierarchical models for animal populations, in a way that separates the processes affecting detectability given the population size, the population when inhabitable, suitable habitat given environmental and climate characteristics, and so on. In psychology, around 28% of Bayesian regression analyses over a quarter of a decade were motivated to use Bayesian to enable computation, or to improve accuracy over Frequentist alternatives (van de Schoot et al., 2017). Advances in Bayesian computation have enabled solution of previously intractable problems, in a wide range of contexts such as: the sudden polarisation of magnets (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953), analysis of medical imagery (Besag, 1986), search and detection of missing aircraft (Stone, Keller, Kratzke, Strumpfer, et al., 2014), estimating the number of species relying on coral reefs (Fisher, O'Leary, Low-Choy, Mengersen, Knowlton, Brainard, and Caley, 2015), meta-analysis accounting for publication bias and low power (Kim, Belland, and Walker, 2017), accounting for measurement, misclassification and missing values in higher education participation (Goldstein, Browne, and Charlton, 2017).

The results of the Quantitative study could not have been obtained using a typical Frequentist approach, because of perfect predictability for some portions of the data — some teachers' decisions exactly reflected student grades. This property was discussed in the context of an examinee correctly answering all items in a test (Levy, 2016, p376), for a simpler IRT model, which here would have corresponded to evaluating whether a teacher's placement decisions perfectly aligned with grades.

Hence our study falls into the last-named category where alternative methods are 'not developed yet'. Bayesian computation was able to accommodate 'perfect predictability', for at least some hypothetical student cards and/or teachers. This would lead to numerical

issues with chi-squared hypothesis testing (due to cells with zero counts), Analysis of Covariance (ANCOVA) or with a Frequentist analysis (because the underlying mathematics allows probabilities between zero and one, but not exactly zero or one). Practically, this meant it was possible to develop an overarching model for all data, rather than identifying then treating the 'perfectly predictable' situations separately.

In addition, in this case we found that a model that only included stereotyping effects for students or threshold effects for teachers did not perform as well as a model that contained both (Riley and Low-Choy, tted). As noted by Kuo and Sheng (2015), when we wish to evaluate the contribution of teacher-specific and student-specific effects at the same time, a Bayesian approach is mandated, since it is the only option: *'In IRT, simultaneous estimation of item [student] and person [teacher] parameters calls for the need of using fully Bayesian estimation via Markov Chain Monte Carlo...'.* This benefit, of being able to fully address uncertainty, is named as the fourth of six benefits of a Bayesian approach for education research (Levy, 2016).

Moreover, Bayesian computation via Markov Chain Monte Carlo is highly flexible (van de Schoot et al., 2017). In this case it enabled us to easily check different model structures.

7. Bridging via Accumulating Knowledge

Typically, many studies are 'one-off', and focus on analysis of a single dataset, large or small. Over the last few decades, researchers have sought to pool results across multiple data sources, by pooling effect sizes over studies with very similar design using meta-analysis (Cooper, 2015; Schmidt and Hunter, 2015) or by pooling data across disparate studies via integrative data analysis (e.g. Brown, Brincks, Huang, Perrino, Cruden, Pantin, Howe, Young, Beardslee, Montag, et al., 2016; Chan, 2017). As discussed in this section, Bayesian analysis offers an explicit mechanism for combining information, in a sequential way, from one study to the next. The prior encapsulates the current state of knowledge, which has been translated or encoded into the relative plausibility of parameters, such as effect-sizes (Low Choy et al., 2009; O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley, and Rakow, 2006). The results of the Bayesian analysis, called the posterior, represents the updated state of knowledge, in terms of the model parameters such as effect sizes.

As discussed in a previous section (Section 1) all of these bridges are made possible because Bayesian statistical analysis incorporates a prior distribution encapsulating the current state of knowledge (prior to examining the new data). In addition, the prior may also have direct benefit: in explicitly capturing the prior state of knowledge. Bayesian statistics provides an explicit mechanism for acknowledging researcher stance *prior* to analysis of the data. This represents prior knowledge gained through experience or from the literature, and codifies knowledge in the form of statistical distributions describing uncertainty about components of the model.

A key defining feature of Bayesian statistical modelling is that it requires the researcher to specify an *a priori* stance about their uncertainty in the model, expressed in terms of their uncertainty about the parameters. This information is usually encoded into a statistical distribution about the model parameters, called the *prior* distribution (Riley and Low-Choy, tted). It is possible for a researcher to claim complete ignorance, and specify a non-informative prior, following a so-called objective stance (Berger, 2006). This is a safe approach to use when implementing Bayesian methods for the first time in a new context (Spiegelhalter et al., 2004). However, as discussed in McElreath (2016), the non-

informative prior is rarely the 'best' approach, as it is usually implausible that nothing is known about model parameters (Goldstein, 2006). Alternatively, it is possible to characterize the researchers' prior knowledge about the model, either based on previously published empirical evidence (e.g. Hobbs and Hilborn, 2006) or by eliciting their expert judgments in a methodical, repeatable way (O'Hagan et al., 2006). In this case study, the initial analysis used a non-informative prior since no prior information was readily available.

Interestingly, this so-called 'subjective' prior (Goldstein, 2006; Press, 2003) at the heart of Bayesian statistics is consistent with the qualitative premise that researchers ought to acknowledge the bias inherent in their own personal perspective: (Glaser and Strauss, 1967, p3) note that:

> *'Of course, the researcher does not approach reality as a tabula rasa. He (sic) must have a perspective that will help him see relevant data and abstract significant categories from his scrutiny of the data'*

Indeed, the prior can provide an important bridge between the qualitative and quantitative aspects of a mixed methods study (Lilford and Braunholtz, 2003).

Similar in aim to a meta-analysis, any Bayesian analysis makes it possible to combine information from different sources. For our initial analysis, we presumed *a priori* a stance of complete ignorance, expressed in what is called a *vaguely informative* prior. If an informative prior is used, then the results end up being essentially a compromise between the prior and the data, as illustrated in Meyer (1964) and van de Schoot et al. (2017) for education audiences. This can be extremely important (O'Hagan et al., 2006; van de Schoot et al., 2017) for example, when the researcher wishes to:

- update an analysis of a moderately large dataset (as discussed here) with new information, with a relatively small sample size;
- provide real-time updating as new streams of information arrive;
- account for weaknesses in empirical data, by incorporating another source of information, for instance obtained in a similar context;
- incorporate carefully elicited expert knowledge, particularly for initial analysis of a small dataset, which would otherwise provide very vague results.

As noted early the use of a vaguely informative prior can in fact be helpful in overcoming numerical issues with fitting the model, for instance involving small data or other complex models.

The results of this first analysis, which did not benefit from any prior information, may provide an informative prior suitable for inclusion for a followup study. Thus a new study could take this study's credible interval as a starting point (the prior): the effect of male gender adjusts the mental score (for females), for a Medium-High grade student, on average by -1 point, with a standard deviation of 1 point. Statistically this can be encoded by a Normal (or white noise) distribution $N(-1, 1)$. The data in this new study could be small (even just a single additional teacher assessing all student cards, or one more student card assessed by all teachers) or large (comprising millions of teachers). The posterior estimate of the range of adjustments would move closer to zero (e.g. $N(0, 2)$ if males in this grade are more similar to females, but could also reduce if males score much lower than females, on average, e.g. $N(-5, 1.5)$. The amount of influence that the new data has is determined by the amount of data, as well as how clear the effects are.

8. Mixing-in Bayesian statistical modelling with Qualitative method: some philosophical implications

At the nexus between quantitative and qualitative methods, it is commonplace to encounter a divergence or even a clash of purist research cultures (Hesse-Biber, 2015, pxxxv), building barriers both conceptually and communication-wise. Alternatively, as presented here, it is possible to identify and then exploit the commonalities under the umbrella of a mixed-in research method. Here we have described, using a case study, five ways that Bayesian analysis can be used whilst bridging this research culture gap. Two of these bridging points arise because Bayesian analysis, like Frequentist approaches, takes a statistical modelling approach. This *broad paradigm of statistical modelling has been replacing the Null Hypothesis Significance Testing (NHST) paradigm* for many decades, with calls for 'Moving from tests to estimates' (Greenland et al., 2016, p340), with debates recently reignited when p-values were outlawed by a social psychology journal (Trafimow and Marks, 2015). Our work demonstrates that this transition, from testing to modelling, is achievable (Rodgers, 2010). The three remaining bridging points are only possible within the Bayesian paradigm. Adopting such a mixed methods approach, which mixes in Bayesian statistics, requires a 'stepchange' in the way that researchers interpret what statistical analysis is, and thus how it can be mixed-in with qualitative research, what this can achieve (and hence how it can be valuable), how such an approach can better address research questions, and how this is achieved. In this section we address the ontological, epistemological, axiological and methodological issues of mixing-in Bayesian with Qualitative method.

8.1. A common ground for Quant/Qual

It has been argued that quantitative and qualitative research need not be differentiated; and in fact 'it can be argued that virtually every study represents a multiple [quant/qual] methods research study to some degree.' (Onwuegbuzie and Hitchcock, 2015, p276). Certainly, this view is supported by the first bridge that constructs a common conceptual and theoretical framework that, in this study, has underpinned both qualitative and quantitative analyses.

These variables are not only crucial to structuring a quantitative analysis, they may also reflect or embody concepts and themes underlying a qualitative analysis. Indeed, some consistency amongst these concepts can aid in integration of qualitative and quantitative components (Maxwell et al., 2015, in final Learnings, p23)

> *... integration is substantially influenced by the mental models used by the researcher(s). Greene (2007) strongly urged that researchers integrate multiple mental models in their work, seeing this as the most developed and enlightening form of mixed methods research. A researcher's inability to understand or appreciate the perspective of someone with a different approach to research can seriously inhibit the integration of different approaches, and "ontological divides" is one of the barriers to integration that Bryman (2007, pp. 16–17) identifies. ... it seems very likely to us that many of these researchers were simply unaware of how their ontological and epistemological assumptions were shaping their research and possibly constraining their integration of their data.*

Thus the creation of a graphical model to support Bayesian analysis may assist researchers cross this 'ontological divide', separating quantitative and qualitative spheres.

Hence this first bridge, which frames the research question within some theoretical framework, may remind researchers of this important foundation for their research. By explaining how to make sense of the research, *this theoretical structure common to qual/quant*

research provides a solid structure for transmitting the epistemology of the research, and greatly facilitates the integration across method.

8.2. Redefining what Quantitative method is

As addressed in the third bridge of Interpretation, Bayesian radically affects the logic of statistical inference, and in fact can make it easier to understand. Interestingly, because of the way in which it is structured, the NHST paradigm leads researchers to concentrate on dichotomous research hypotheses, which either over-simplifies the underlying theory, or over-complicates these, in ways which can be very confusing[4], and do not invite engagement or conversation. Indeed the logic underlying NHST is prone to severe misinterpretation (e.g. as summarized in Greenland et al., 2016) and exemplifies the 'Inversion Fallacy' known to philosophers since the fifties (Villejoubert and Mandel, 2002). This widespread misinterpretation of what statistical analysis is (ontologically), and its logical underpinnings (epistemology), can lead to misleading findings and non-reproducible research (e.g. Etz and Vandekerckhove, 2016), which has negatively affected its perceived value (axiology), leading to distrust in statistics as an approach, and as a cornerstone of reproducible research (Huff, 1993).

In contrast, the popularity of graphical models (Düspohl, Frank, and Döll, 2012; Greenland and Brumback, 2002) indicates that researchers find it useful to harness visual means to communicate—to peers and endusers—the complexities of the underlying theory. The research questions then emerge from, and are tied to, this theory. This naturally encourages a method of multiple, rather than single, working hypotheses, as highlighted four decades before NHST by a geologist (Chamberlin, 1890). Indeed the Bayesian paradigm explicitly evaluates the plausibility of a continuum of hypotheses, and is not restricted to single overly-specific hypotheses, such as the nil (null) hypothesis or straw man. In this way, *this redefines how quantitative analysis may be enacted (epistemologically). Instead of a 'one-off' rules-based approach (Gigerenzer et al., 2004), statistical modelling through a Bayesian lens can be defined, ontologically, to involve an iterative though principled and logical approach*, e.g. within the hypothetico-deductive paradigm (Gelman and Shalizi, 2013). This redefines (axiologically) what can be achieved through quantitative analysis, in terms of a broader response to research questions.

8.3. Communication between Quant/Qual

The third bridge of Interpretation, through Stories enables researchers to remain faithful to the individual stories contained in the individuals responding to both qualitative and quantitative aspects of the study. This is related to 'process tracing' (Humphreys and Jacobs, 2015, p656). Placing the stories at the centre of interpretation, under both qual/quant research methods, is a form of 'drilling-down' from the analysis of patterns and trends into the raw data. In the fields of data visualization and precision journalism, this technique is well-respected as a means of communicating complexity and engendering credibility (Tufte, 1983). Story-telling is able to take advantage of the *the increased complexity available within the Bayesian paradigm when posing research questions, and subsequent interpretation of answers*. As noted by Ziliak and McCloskey (2008, p151), in practice it can be difficult to falsify a simple hypothesis, when it is appropriately considered in context: 'scientific hypotheses are accompanied by side conditions—instrumentation or controls— making

[4] The American Statistical Association has recently formulated guidelines on how to interpret p-values which have been embedded in statistical teaching since the sixties (Wasserstein and Lazar, 2016)

the observation possible'. Thus story-telling has the potential for making contextualization more explicit, and hence reduce the conflict between the highly nuanced specificity inherent in qualitative research versus the generalizations required to meet the demands of a quantitative approach (Onwuegbuzie and Hitchcock, 2015, p275).

8.4. Redefining what can be achieved via Quantitative methods

Methodologically, computation underlying the Bayesian approach may have a very different mathematical 'flavour', involving simulation as well as numerical approximations, with less reliance on large sample assumptions, compared to Frequentist counterparts (Jackman, 2009, Introduction on 'Why be Bayesian?'). *One advantage of Bayesian computation is that it can enable investigation of models that cannot be fit within the Frequentist or NHST paradigms*, as considered here under the fourth point of bridging. Again, this subtly moves the boundaries delineating what can be achieved via statistical analysis, and hence revises its ontology and axiology.

Furthermore, successful Bayesian software and languages (such as Spiegelhalter et al., 2003, , a seminal WinBUGS package) provide a 'modelling language' that allows the modeller to 'plug-and-play' with modelling components, thereby creating a 'bespoke' model that is tailored to the problem at hand. Here the Item Response Theory model was crafted to include components standard to a Rasch model (corresponding to the teacher's mental score for each hypothetical student), and extend it in various ways, including components to allow for each teacher's style in choosing thresholds on these scores. Because of the flexible construction of Bayesian statistical models, each element can be adjusted, which greatly facilitates comparison amongst models, and necessitates a more iterative approach to modelling (e.g. Gelman and Shalizi, 2013).

8.5. The potential for Bayesian in Mixing-in of Methods

In the context of education (and indeed the social sciences more broadly), we are beginning to explore a Bayesian approach to mixing-in of methods, for application to problems where it is well suited, e.g. for:

- *small studies* that are easier to implement, such as course analytics in higher education where feedback from students is both quantitative (such as surveys with Likert-scale responses) as well as qualitative (e.g. focus groups) in order to help revise curriculum, techniques and/or tools;
- investigations of *complex systems* (such as evaluation of educational systems and/or interventions, or learning analytics) that have access to *large datasets* (which only tell part of the story in a simplified kind of way, for a large number of individual cases) and thus would benefit from supplementation with *interviews*, that tell a more detailed story for a smaller number of cases;
- researchers who have been advised to 'do more than a multiple regression' and consider a *hierarchical model*, via a technique such as Structural Equation Modelling (SEM), but find that it is difficult for various reasons, such as:
    - it is challenging to get started, to discern the link between their conceptual knowledge and the model structure,
    - their problem doesn't easily fit into standard templates (e.g. multiple issues such as missing values, non-normality, excess zeros, nested relationships),
    - they wish to integrate expert knowledge or findings from previous studies in a transparent way, that is also easily accessible to a wide audience—for instance,

Humphreys and Jacobs (2015) provide an explicit technical framework for integrating particular kinds of information using a Bayesian Integrated Quantitative & Qualitative framework;
    - they find it difficult to interpret and communicate the model findings (what do the statistical tests mean for an SEM?);
- when a sequential approach to learning is required, to enable feasible implementation over a longer period of time, which is flexible to changes in both quantitative and qualitative information sources and collaborative arrangements: as described in 'The Bayesian Superintendent' by Meyer (1964) and more generally in education by König and van de Schoot (2017);

We note that in several arenas, Bayesian methods are being exploited for their computational benefits, but analysis and reporting may still being 'shoehorned' into the same framework as Frequentist methods, e.g. by simply replacing confidence intervals for credible intervals. Whilst this can be a staged way of enabling a transition, it means that the analysis has not reached its full potential, as afforded by the Bayesian paradigm. For instance, the observation that 'It could be that psychologists do not readily identify situations in which the use of Bayesian methods may be beneficial.' (van de Schoot et al., 2017) could apply to researchers in education as well. In particular this often precludes the possibility of taking advantage of Bayesian benefits when mixing-in with qualitative research methods.

8.6. Redefining how Quantitative analysis is done

Because of the underlying computation, it is rarely possible to conduct Bayesian analysis 'on-the-back-of-an-envelope', in the same way that many NHST techniques can. Instead, it is more important that modellers and researchers using the model understand the mathematics in the model; 'A statistical model is much more than an equation with Greek letters' (Greenland et al., 2016, p345). In our case study, as facilitated by the second and third bridges, this transition requires (simpler) mathematics to understand the model rather than the calculations. Again this redefines the way in which qualitative researchers can approach quantitative analysis, from the basis of the conceptual model (the first point of bridging) rather than as a vehicle for improving their mathematical skills. Thus, 'The epistemological basis of statistics has moved away from being a set of procedures, applied mechanistically, and moved toward building and evaluating statistical and scientific models' (Rodgers, 2010).

The transparent statement of prior beliefs underlie a fifth point of bridging between quant/qual (Sections 1, 7.) The prior is at the heart of the mathematical and epistemological difference between Frequentist and Bayesian approaches to statistical modelling. This instigates a change in axiology: by redefining the value of other sources of information (previous studies or expert knowledge) encapsulated in the prior model (Low Choy et al., 2009). In turn, this facilitates a change in focus from a search for an immutable truth, into a comparison of prior and posterior models, and hence a question of how the most recent empirical data has modified our understandings (Lilford and Braunholtz, 2003). This is the fifth point of bridging addressed in Section 7. Bayesian updating helps address the new emphasis on reproducible research (e.g. Etz and Vandekerckhove, 2016), and supports interpretation of statistical enquiry as a process that extends across multiple studies. Importantly, for mixed methods, prior studies can be qualitative (as in Section 7) or quantitative: the main challenge is to quantify their findings in a way that is consistent within the most recent conceptual framework and hence statistical model. Indeed,

Humphreys and Jacobs (2015) develop an explicit framework, the Bayesian integration of quantitative and qualitative data (BIQQ), that specifies how a certain form of qualitative evidence (expressed in terms of clues) can be combined with quantitative data, to provide stronger models of causality that combine both sources of information.

9. Conclusions

Every year, online technologies are providing an avalanche of richer, more complex information in greater quantities. For education researchers well versed in qualitative research and traditional statistical analysis, the process of making sense of this information may encounter severe technical and intellectual obstacles. Mixed methods has the potential to integrate the research findings across the quantitative and qualitative spheres. We assert that the statistical modelling paradigm chosen for the quantitative component can influence how mixing-in is achieved. Here we identify five ways in which Bayesian statistical modelling can form 'bridges' with the qualitative component of the study. We note whether bridges are afforded by a model-based approach (which may be either Frequentist or Bayesian), or whether they are particular to the Bayesian approach. Because we are focussed on Bayesian bridges to Qualitative method, our perspective here is necessarily anchored on the quantitative component.

To illustrate how bridging can be achieved whilst mixing-in Bayesian with qualitative methods, we use a case study investigating how teachers' decisions were influenced by the students' grades and the teachers' stereotyping. Mixed methods were used to link the results of the behavioural study—which asked teachers to make decisions to allocate hypothetical students to different levels of assistance—with the insights provided by thematic analysis of probing follow-up questions (perceptions regarding their decisions) as well as interviews (again perceptions regarding the issues in general). By mixing in the Behavioural, Perceptual and Interview questions, the principal investigator obtained a rich, multi-faceted dataset. However, similar to many problems in the social sciences, the nature of the data required quite sophisticated analysis, that falls what is currently easily accessible in standard statistical packages, and necessitated a bespoke Bayesian model.

The first of the five bridges is achieved via *visualization* of the underlying theories and conceptual framework, which enables a smooth transition to later construction and *interpretation* of the Bayesian statistical model (the third bridge). We progress from a Venn diagram that delineates the relevant sectors of literature, to a mind-map and then an influence diagram of concepts and their inter-relationships, culminating in a Directed Acyclic Graph that expresses the variables and dependencies contained in the Bayesian statistical model. In this way, the visual representations gradually refine the way in which the theories (on expectations, attributions and stereotyping) support and increasingly target the research questions regarding how teachers' decision-making could be influenced by stereotyping and other factors. More generally, we note that exploiting visualization is potentially one of the most powerful devices for crossing the disciplinary divide between qualitative and quantitative researchers, as evidenced by current trends in social and behavioural research to rely more and more on graphical models, such as Structural Equation Models and Bayesian networks.

Interview components this study was able to produce data and subsequent analyses that satisfied the contextualized specificity demanded by qualitative methods, simultaneously with the precise characterization of less specific features, that enable quantitative analysis to be generalized beyond the data examined. These *generalizations* were in fact made possible through a third bridge, via Bayesian notions of exchangeability and the flexible

hierarchical structure, replacing the obligations of full randomization that proved infeasible in this context. Self-selecting teachers could be considered exchangeable amongst engaged teachers in the broader population, but would not be considered a random sample desirable for Frequentist analysis. Due to constraints with the online tool, the factors of interest (student grades and triggers for teacher stereotyping) were randomized once for all teachers. Bayesian analysis could predicate analysis on this single ordering of cards, whereas this violated the classical requirements of a Frequentist analysis. This kind of conditional analysis makes use of the flexible framework of Bayesian thinking, which is yet to be fully appreciated, given the pre-existing strictures for interpreting NHST.

As demonstrated here, a Bayesian statistical approach can enhance a mixed methods study, because of its ability to explore and highlight the nuances that are inherent in: the research questions, the overall purpose (and hence desired generalizations), the data and its design, and importantly, for communication with qualitative researchers, the *interpretation* of the findings. Equally, qualitative research has the ability to enhance Bayesian analyses, by exemplifying the general patterns and trends using real individuals. Whilst this bridge can be constructed for a Frequentist analysis, the flexibility to consider individualized random effects with uncertainty, which is only available through the Bayesian paradigm, makes it much easier to trace individual cases. Such nuances are particularly important in an online setting, where interactions are less likely to be in person. Interestingly, extensive research on capturing expert knowledge and representing the current state of knowledge (in priors) provides a foundation for dealing with these nuances.

We note that traditional confirmatory analyses (within the NHST or Frequentist paradigms) still have their place in some research studies, when there is a single clear null hypothesis to be evaluated. However, in many cases researchers wish to establish which hypotheses (and models) are more plausible. Hence educational research is not predominantly aimed towards confirmation and verification. Instead, much research in this field is exploratory or pioneering, particularly in the rapidly expanding digital realm.

Given the escalating amount of online learning, education and research, it will be more important that researchers in qual/quant methods can work together in more meaningful ways. We are currently experiencing a formative period for mixed methods, in all disciplines. It is particularly relevant to education, where traditionally the 'art' of teaching has relied on multiple sources of information about their efficacy. However, teachers and education systems are under pressure due to the explosion of 'big data' in this realm. These bridges can help integrate diverse sources of information, engage with researchers transitioning from purely qualitative to mixed methods approaches, and engender communication between the diverse qual/quant cultures. By building and communicating via the bridges provided by a Bayesian approach (including the five illustrated here) we obtain more opportunity for building interpersonal relationships: for seeding conversations about the shared research problem, for building a shared language about the research, for sharing multiple perspectives, and also revealing points of difference. Together these allow moments for real collaboration and communication among quant/qual researchers.

Acknowledgement

The ideas behind this paper have been nurtured and sharpened through the accumulation of many small yet stimulating dialogues. The authors gratefully acknowledge: (a) the Special Interest Group (SIG) in the Sociology of Education (SocEd), led by Parlo Singh, and in particular: Chris Bigum, Katherine Main, Stephen Heimans, Ben Williams, Sue Whatman and Judy Rose; (b) "the Bayesian Exchange", a reading group of statistical researchers at

Griffith University, especially Daniela Vasco, Ramethaa Pirathiban, Cameron Williams, Ibrahim Al-Khairy, Zarina Vahitkova and Ben Stewart-Koster; (c) the other members of the Curious Collective in the Griffith Institute of Educational Research: Naomi Barnes, Sherilyn Lennon, and Sue Monk. We also thank an Anonymous Reviewer for very helpful guidance and comments.

References

Bayarri, M. J. and J. O. Berger (2004, 02). The interplay of Bayesian and Frequentist analysis. *Statist. Sci. 19*(1), 58–80.

Bazeley, P. and K. Jackson (2013). *Qualitative data analysis with NVivo* (2nd ed.). London: Sage.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis 1*, 385–402.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological) 48*(3), 259–302.

Bijak, J. and J. Bryant (2016). Bayesian demography 250 years after Bayes. *Population Studies 70*(1), 1–19.

Box, G. E. P., J. S. Hunter, and W. G. Hunter (2005). *Statistics for experimenters: Design, Innovation, and Discovery* (Second ed.). Wiley Series in Probability and Statistics. Wiley: Hoboken, NJ.

Brown, C. H., A. Brincks, S. Huang, T. Perrino, G. Cruden, H. Pantin, G. Howe, J. F. Young, W. Beardslee, S. Montag, et al. (2016). Two-year impact of prevention programs on adolescent depression: an integrative data analysis approach. *Prevention Science*, 1–21.

Browne, W. J. and D. Draper (2006, 09). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis 1*(3), 473–514.

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research 1*(1), 8–22.

Campitelli, G. and G. Macbeth (2014). Hierarchical graphical bayesian models in psychology. *Revista Colombiana de Estadística 37*(2), 319–339.

Chamberlin, T. C. (1890). The method of multiple working hypotheses. *Science 15*, 92?96. (reprinted in Science 148: 754?759 [1965]).

Chan, M. L. (2017). An explicit pragmatic approach to integrative data analysis strategies for mixed methods research. *International Journal of Linguistics 9*(3), 166–184.

Cohen, S. (2016). *Bayesian Analysis in Natural Language Processing*. Morgan & Claypool: California.

Cooper, H. (2015). *Research synthesis and meta-analysis: A step-by-step approach*, Volume 2. Sage publications.

Cooper, T. J., A. R. Baturo, E. Warren, and S. M. Doig (2004). Young white teachers' perceptions of mathematics learning of aboriginal and nonaboriginal students in remote communities. In M. J. Hoines, A. Fuglestad, and A. Berit (Eds.), *28th Annual Conference of the International Group for the Psychology of Mathematics Education (PME), July 14-18, 2004, Bergen, Norway*. International Group for the Psychology of Mathematics Education, Cape Town. Available from: `http://igpme.org`.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

Dandy, J., K. Durkin, B. Barber, and S. Houghton (2015). Academic expectations of Australian students from Aboriginal, Asian and Anglo backgrounds: Perspectives of teachers, trainee-teachers and students. *International Journal of Disability, Development and Education 62*(1), 60–82.

De Boeck, P., M. Bakker, R. Zwitser, M. Nivard, A. Hofman, F. Tuerlinckx, I. Partchev, et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software 39*(12), 1–28.

Diamond, J. (1964). Bayesian statistics: a place in educational research? Available from `http://eric.ed.gov/?id=ED069707`.

Drabble, S. J. and A. O'Cathain (2015). Moving from randomized controlled trials to mixed methods intervention evaluations. In S. N. Hesse-Biber and R. B. Johnson (Eds.), *The Oxford Handbook for multimethod and mixed methods research inquiry*. Oxford University Press.

Draper, D., J. S. Hodges, C. L. Mallows, and D. Pregibon (1993). Exchangeability and data analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 156*(1), 9–37.

Düspohl, M., S. Frank, and P. Döll (2012). A review of bayesian networks as a participatory modeling approach in support of sustainable environmental management. *Journal of Sustainable Development 5*(12), 1.

Elliott, L. P. and B. W. Brook (2007). Revisiting chamberlin: Multiple working hypotheses for the 21st century. *BioScience 57*(7), 608–614.

Ellison, A. M. (1996). An introduction to Bayesian inference for ecological research and environmental decision making. *Ecological Applications 6*, 1036–1046.

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters 7*, 509–520.

Etz, A. and J. Vandekerckhove (2016). A bayesian perspective on the reproducibility project: Psychology. *PLoS ONE 11*(2), e0149794.

Fidler, F. and G. Cumming (2005). Teaching confidence intervals: Problems and potential solutions. In *International Statistical Institute, 55th Session*.

Finch, W. H. and B. F. French (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods 11*(1), 14.

Fisher, R., R. A. O'Leary, S. Low-Choy, K. Mengersen, N. Knowlton, R. E. Brainard, and M. J. Caley (2015). Species richness on coral reefs and the pursuit of convergent global estimates. *Current Biology 25*(4), 500 – 505.

Frigg, R. and S. Hartmann (2017). Models in science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 ed.). Metaphysics Research Lab, Stanford University.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*(3), 515–533.

Gelman, A. and C. R. Shalizi (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology 66*, 8–38.

Gigerenzer, G., S. Krauss, and O. Vitouch (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, pp. 392–409. SAGE Publications, Inc.

Glaser, B. and A. Strauss (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine.

Goldstein, H., W. J. Browne, and C. Charlton (2017). A bayesian model for measurement and misclassification errors alongside missing data, with an application to higher education participation in australia. *Journal of Applied Statistics 0*(0), 1–14.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis 1*, 403–420.

Greenland, S. and B. Brumback (2002). An overview of relations among causal modelling methods. *International journal of epidemiology 31*(5), 1030–1037.

Greenland, S., R. Daniel, and N. Pearce (2016). Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *International journal of epidemiology 5*(2), 565–57.

Hesse-Biber, S. (2015). Introduction: Navigating a turbulent research landscape: Working the boundaries, tensions, diversity, and contradictions of multimethod and mixed methods inquiry. In S. N. Hesse-Biber and R. B. Johnson (Eds.), *The Oxford Handbook for multimethod and mixed methods research inquiry*, pp. xxxiv–liii. Oxford University Press.

Hesse-Biber, S., D. Rodriguez, and N. A. Frost (2015). A qualitatively driven approach to multimethod and mixed methods research. In S. N. Hesse-Biber and R. B. Johnson (Eds.), *The Oxford Handbook for multimethod and mixed methods research inquiry*, Chapter 1, pp. 3–20. Oxford University Press.

Hilborn, R. and M. Mangel (1997). *The Ecological Detective: Confronting models with Data*. Monographs in Population Biology, 28. Princeton University Press: Princeton, New Jersey.

Hobbs, N. T. and R. Hilborn (2006). Alternatives to statistical hypothesis testing in ecology: A guide to self teaching. *Ecological Applications 16*, 5–19.

Huff, D. (1993). *How to lie with statistics*. New York: Norton.

Hulting, F. L. and D. A. Harville (1991). Some Bayesian and non-Bayesian procedures for the analy-

sis of comparative experiments and for small-area estimation: Computational aspects, Frequentist properties, and relationships. *Journal of the American Statistical Association 86*(415), 557–568.

Humphreys, M. and A. M. Jacobs (2015). Mixing methods: A Bayesian approach. *American Political Science Review 109*(4), 653–673.

Jackman, S. (2009). *Bayesian analysis for the social sciences*, Volume 846 of *Wiley Series in Statistics and Probability*. John Wiley & Sons: Chichester, UK.

Kelle, U. (2015). Mixed methods and the problems of theory building and theory testing in the social sciences. In S. N. Hesse-Biber and R. B. Johnson (Eds.), *The Oxford Handbook for multimethod and mixed methods research inquiry*. Oxford University Press.

Kieftenbeld, V., P. Natesan, and C. Eddy (2011). An item response theory analysis of the mathematics teaching efficacy beliefs instrument. *Journal of Psychoeducational Assessment 29*(5), 443–454.

Kim, N. J., B. R. Belland, and A. E. Walker (2017, Jul). Effectiveness of computer-based scaffolding in the context of problem-based learning for STEM education: Bayesian meta-analysis. *Educational Psychology Review*. Available from: `https://doi.org/10.1007/s10648-017-9419-1`.

König, C. and R. van de Schoot (2017). Bayesian statistics in educational research: a look at the current state of affairs. *Educational Review 0*(0), 1–24.

Kuo, T.-C. and Y. Sheng (2015). Bayesian estimation of a multi-unidimensional graded response IRT model. *Behaviormetrika 42*(2), 79–94.

Levy, R. (2016). Advances in Bayesian modeling in educational research. *Educational Psychologist 51*(3-4), 368–380.

Lilford, R. J. and D. Braunholtz (2003). Reconciling the Quantitative and Qualitative traditions: The Bayesian approach. *Public Money & Management 23*(3), 203–208.

Lindley, D. V. and M. R. Novick (1981). The role of exchangeability in inference. *Journal of the American Statistical Association 9*, 45–58.

Low-Choy, S. (2013). Priors: Silent or active partners in Bayesian inference? In A. C., K. Mengersen, and A. N. Pettitt (Eds.), *Case Studies in Bayesian Statistical Modelling and Analysis*, pp. 30–65. John Wiley & Sons, Inc, London.

Low Choy, S., R. O'Leary, and K. Mengersen (2009). Elicitation by design for ecology: using expert opinion to inform priors for Bayesian statistical models. *Ecology 90*, 265–277.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing 10*(4), 325–337.

Ly, A. and L. Crowshoe (2015). Stereotypes are reality: addressing stereotyping in Canadian Aboriginal medical education. *Medical education 49*(6), 612–622.

Mark, M. M. (2015). Mixed and multimethods in predominantly quantitative studies, especially experiments and quasi-experiments. In S. N. Hesse-Biber and R. B. Johnson (Eds.), *The Oxford Handbook for multimethod and mixed methods research inquiry*, Chapter 2, pp. 21–41. Oxford University Press.

Maxwell, J. A., M. Chmiel, and S. E. Rogers (2015). Designing integration in multimethod and mixed methods research. In S. N. Hesse-Biber and R. B. Johnson (Eds.), *The Oxford Handbook for multimethod and mixed methods research inquiry*, Chapter 13, pp. 223–239. Oxford University Press.

McCambridge, J., K. Kypri, and D. Elbourne (2014). In randomization we trust? there are overlooked problems in experimenting with people in behavioral intervention trials. *Journal of Clinical Epidemiology 67*(3), 247 – 253.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*, Volume 122. Boca Raton, FL: Chapman & Hall/CRC.

Merriem, S. and E. Tisdell (2015). *Qualitative research: A guide to design and implementation* (fourth ed.). San Francisco: Jossey-Bass.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Meyer, D. L. (1964). A Bayesian school superintendent. *American Educational Research Journal 1*(4), 219–228.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika 59*(4), 439–483.

Moynihan, C., R. Lewis, E. Hall, E. Jones, A. Birtle, and R. Huddart (2012, Nov). The patient deficit model overturned: a qualitative study of patients' perceptions of invitation to participate in a randomized controlled trial comparing selective bladder preservation against surgery in muscle invasive bladder cancer (spare, cruk/07/011). *Trials 13*(1), 228.

Oakes, J. and G. Guiton (1995). Matchmaking: The dynamics of high school tracking decisions. *American Educational Research Journal 32*(1), 3–33.

O'Hagan, A., C. E. Buck, A. Daneshkhah, R. Eiser, P. Garthwaite, D. Jenkinson, J. Oakley, and T. Rakow (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley.

Oldehinkel, A. J. (2016). Editorial: Bayesian benefits for child psychology and psychiatry researchers. *Journal of Child Psychology and Psychiatry 57*(9), 985–987.

Onwuegbuzie, A. J. and J. H. Hitchcock (2015). Advanced mixed analysis approaches. In S. N. Hesse-Biber and R. B. Johnson (Eds.), *The Oxford Handbook for multimethod and mixed methods research inquiry*, Chapter 16, pp. 275–295. Oxford University Press.

Pickering, C. and J. Byrne (2014). The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers. *Higher Education Research & Development 33*(3), 534–548.

Pit-ten Cate, I. M., S. Krolak-Schwerdt, and S. Glock (2016). Accuracy of teachers' tracking decisions: Short- and long-term effects of accountability. *European Journal of Psychology of Education 31*(2), 225–43.

Press, S. J. (2003). *Subjective and objective Bayesian statistics: principles, models, and applications*. Wiley: New Jersey. Second Edition.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rainey, C. (2016). Dealing with separation in logistic regression models. *Political Analysis 24*(3), 339–355.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhangen: Danmarks Paedogogiske Institut.

Raudenbush, S. W., A. S. Bryk, Y. F. Cheong, R. Congdon, and M. du Toit (2004). *HLM 6: Hierarchical Linear and Nonlinear Modeling*. Scientific Software International, Lincolnwood, IL. Available from: `http://www.ssicentral.com/hlm/`.

Riley, T. (accepted). Exceeding expectations: teacher's decision making regarding Aboriginal and Torres Strait Islander students. *Journal of Teaching & Education 0*. Accepted for publication subject to minor revisions, 9 May 2017.

Riley, T. and S. Low-Choy (submitted). Same grades, different placement: Linking a Bayesian ordinal regression with interview data to reveal the influence of teachers' stereotypes. Submitted for publication.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *American Psychologist 65*(1), 1–12.

Royle, J. A. and R. M. Dorazio (2008). *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press, Elsevier, London.

Schmidt, F. L. and J. E. Hunter (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (third ed.). Sage publications: CA.

Sedimeier, P. and G. Gigerenzer (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General 130*(3), 380–400.

Shoemaker, J. S., I. S. Painter, and B. S. Weir (1999). Bayesian statistics in genetics: a guide for the uninitiated. *Trends in Genetics 15*(9), 354–358.

Smith, B. J. (2007). boa: An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software 21*(11), 1–37.

Spiegelhalter, D. J., K. R. Adams, and J. P. Myles (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Statistics in Practice. Chichester, U. K.: John Wiley & Sons Ltd.

Spiegelhalter, D. J., A. Thomas, N. G. Best, and D. Lunn (2003). WinBUGS version 1.4 user manual. Technical report, MRC Biostatistics Unit, Cambridge.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of Frequentist and Bayesian approaches. *American Journal of Political Science 57*(3), 748–761.

Stone, L. D., C. M. Keller, T. M. Kratzke, J. P. Strumpfer, et al. (2014). Search for the wreckage of Air France Flight AF 447. *Statistical Science 29*(1), 69–80.

Suess, E. A., I. A. Gardner, and W. O. Johnson (2002). Hierarchical Bayesian model for prevalence inferences and determination of a country's status for an animal pathogen. *Preventive Veterinary Medicine 55*, 155–171.

TechSmith (2016). Camtasia help. Technical Report Version 9.1, September 2017, TechSmith Corporation. Available from: https://support.techsmith.com/hc/en-us/article_attachments/115002225672/Camtasia_9.1_Help.pdf.

Trafimow, D. and M. Marks (2015). Editorial. *Basic and Applied Social Psychology 37*(1), 1–2. DOI: 10.1080/01973533.2015.1012991.

Tufte, E. (1983). *The Visual Display of Quantitative Information*. Graphics Press: Cheshire USA.

van de Schoot, R., S. D. Winter, O. Ryan, M. Zondervan-Zwijnenburg, and S. Depaoli (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods 22*(2), 217.

van Someren, M. W., Y. F. Barnard, and J. A. C. Sandberg (1994). *The think aloud method: A practical approach to modelling cognitive processes (Knowledge-based systems)*. London: Academic Press.

Villejoubert, G. and D. R. Mandel (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory and Cognition 30*(2), 171–178.

Waller, N. G. and W. O. Johnson (1998). The non-significance of straw man arguments. *Behavioral and Brain Sciences 21*(2), 226?227.

Wasserstein, R. L. and N. A. Lazar (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician 70*(2), 129–133. DOI:10.1080/00031305.2016.1154108.

Weiner, B. (1974). *Achievement motivation and attribution theory*. General Learning Press.

Western, B. and S. Jackman (1994). Bayesian inference for comparative research. *American Political Science Review 88*(02), 412–423.

Ziliak, S. T. and D. N. McCloskey (2008). *The Psychology of Psychological Significance Testing*, Chapter 13, pp. 140–153. University of Michigan Press.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis 13*(2), 157–170.