Received: 25 May 2021 | Accepted: 3 May 2022

DOI: 10.1111/bjop.12577

ARTICLE

Statistical modelling of vignette data in psychology

Thom Baguley | **Grace Dunham** | **Oonagh Steer**

Nottingham Trent University, Nottingham, UK

Correspondence
Thom Baguley, Department of Psychology, Nottingham Trent University, 50 Shakespeare Street, Nottingham, NG1 4FQ, UK.
Email: thomas.baguley@ntu.ac.uk

Abstract

Vignette methods are widely used in psychology and the social sciences to obtain responses to multi-dimensional scenarios or situations. Where quantitative data are collected this presents challenges to the selection of an appropriate statistical model. This depends on subtle details of the design and allocation of vignettes to participants. A key distinction is between factorial survey experiments where each participant receives a different allocation of vignettes from the full universe of possible vignettes and experimental vignette studies where this restriction is relaxed. The former leads to nested designs with a single random factor and the latter to designs with two crossed random factors. In addition, the allocation of vignettes to participants may lead to fractional or unbalanced designs and a consequent loss of efficiency or aliasing of the effects of interest. Many vignette studies (including some factorial survey experiments) include unmodeled heterogeneity between vignettes leading to potentially serious problems if traditional regression approaches are adopted. These issues are reviewed and recommendations are made for the efficient design of vignette studies including the allocation of vignettes to participants. Multilevel models are proposed as a general approach to handling nested and crossed designs including unbalanced and fractional designs. This is illustrated with a small vignette data set looking at judgements of online and offline bullying and harassment.

KEYWORDS

factorial survey experiments, multilevel modeling, vignette data

BACKGROUND

Vignette methods are widely used in psychology and the social sciences. Originally, a blend of experimental and survey methodology developed for quantitative measurement of social judgements in sociology (Rossi & Nock, 1982), and vignette methods have been adapted for application in a range of disciplines. Despite this, there is a lack of consensus on how best to model quantitative vignette data (see Wallander, 2009). The present paper outlines key characteristics of vignette studies and considers how the design of the study and vignettes impact the appropriate statistical model.

Vignettes are multidimensional stimuli consisting of 'a short, carefully constructed description of a person, object or situation, representing a systematic combination of characteristics' (Atzmüller & Steiner, 2010, p. 128). Consider the following vignette adapted from Rossi and Anderson, 1982:

> Cindy M., a married graduate student often had occasion to talk to Gary T., a single 65-year old professor. They were both at a party. She said that she enjoyed and looked forward to his class. He asked her about her other courses. He said that she could substantially improve her grade if she cooperated.

The vignette was generated by combining levels on eight different dimensions.

Participants then rated the vignette on a scale of 1 'definitely not harassment' to 9 'definitely harassment'. By varying the dimensions in a systematic way, the study can explore contextual and other factors that influence harassment judgements. For example, the male character's physical behaviour is drawn from ten options ranging from blank text (as above) to overt acts such as 'He put his arm on her shoulder'. A quirk of this example is that some elements (here the names of the characters) vary between vignettes but are not dimensions in the factorial design. This detail turns out to have potentially important implications for the appropriate statistical model. One advantage of the vignette approach is the sheer versatility of the method, and it is increasingly common to use tabular, pictorial or video presentation (Aguinis & Bradley, 2014; Auspurg & Hinz, 2015). Likewise, responses involve a wide variety of formats (though rating scales are particularly common).

The origins of vignette methods in sociology and psychology

Modern vignette approaches date back to Peter Rossi's Ph.D. research on social stratification advised by Paul Lazarsfeld. Lazarsfeld suggested systematically rotating household characteristics via vignettes, using a factorial design (see Rossi & Nock, 1982, p. 9). An important insight was that it was possible to present each respondent with a subset of the *vignette universe* (the superset of possible vignettes). It turns out that the precise construction and allocation of vignettes have important implications for the statistical treatment of vignette data. Work in sociology generally sticks closely to Rossi's original formula for the design and analysis of these studies, though vignettes are used in many other disciplines, and many implementations adopt a more relaxed approach. This may be problematic if quantitative data are analysed using an inappropriate statistical model depending on the aim of the research. In particular, qualitative studies typically do not focus on statistical generalization; either not seeking to generalize at all or considering different forms of generalization (Smith, 2018).

A related method, the situational judgement test (SJT), was developed by psychologists in the 1940s for personnel selection (Whetzel et al., 2020). SJTs elicit responses to situations (e.g., critical incidents) that might arise in a particular role, and participants are typically required to indicate how they would behave in response (Lievens et al., 2008). However, the focus in SJTs is on measurement – including psychometric properties such as reliability and validity (Corstjens et al., 2017; Webster et al., 2020), leading to different statistical considerations when selecting a model. Nevertheless, the SJT literature provides a rich source of advice on constructing realistic scenarios or where measurement is the principal aim.

Vignette and related methods in psychological research

Vignette and closely-related methods have proven particularly popular in quantitative and qualitative psychological research for two main reasons. First, they have the potential to increase the fidelity of stimuli, the extent to which they preserve key properties of the situations of interest, by tailoring both their content and format to the appropriate context or contexts (Finch, 1987; Lanza & Carifio, 1992).[1] This includes enhancing engagement with the research by making the content personally meaningful and relevant or making it easier to interpret. For instance, job applicants seem to view SJTs positively because of the job-relatedness of the content, especially with video, multimedia or interactive content (Chan & Schmitt, 1997; Lievens et al., 2008).

Second, they allow researchers to investigate sensitive topics such as those involving 'intimate, discreditable or incriminating behavior' (de Groot et al., 2020, p. 1), which cannot be assessed directly for ethical reasons. A useful property in this context is that researchers can sidestep the requirement to define crucial terms (e.g., bullying or discrimination) by embedding specific behaviours within the vignettes. This is useful when definitions are contested, have varying interpretations or might bias participants (e.g., explicitly labeling a behaviour as bullying might make an individual less likely to endorse it).

The general adaptability of the approach lends it to working with a wide range of participants and topics. Working with health or medical professionals, vignettes can be adapted to describe patients or case histories. For example, Lewis et al. (1990) presented psychiatrists with case histories varying in their race and gender to see how this impacted assessment or management of the patient. Sauer (2011) used video-presented vignettes to explore the impact of leadership status and style on leader effectiveness. This study also replicated the effects with an in-person experimental study, providing direct evidence that low-to-medium fidelity stimuli can be sufficient to generalize to more realistic contexts. Hine (2019) used vignettes to explore judgements of domestic violence by varying perpetrator and victim gender, and type of abuse (physical, psychological/emotional and financial), illustrating their value in exploring sensitive topics. The flexibility of the method means that such topics can be explored with vulnerable participants if the materials are carefully designed. For example, Maieron et al. (1996) have successfully used vignettes to investigate children's perceptions and acceptance of a hypothetical peer with AIDS.

How the design of a vignette study impacts the choice of analysis

To illustrate the main issues, it is helpful to distinguish between more and less strictly-controlled types of vignette study. At the strictest level are the *factorial survey experiments* pioneered by Rossi and colleagues, in which participants each receive a different allocation of vignettes (in text or tabular form) sampled from the universe of possible vignettes generated by the factorial design (Auspurg & Hinz, 2015; Rossi & Anderson, 1982). This approach allows one to explore the influence of numerous dimensions on participants' responses without presenting all possible vignettes. Furthermore, this approach also either eliminates (or at least minimizes) the impact of collinearity on estimates of the dimension effects. This is achieved by ensuring either that each participant experiences a unique set of vignettes or that overlap is sufficiently rare that confounding is negligible for the effects of interest. In contrast, *experimental vignette studies* (Aguinis & Bradley, 2014) systematically vary the characteristics of the vignettes but relax the requirement to sample from a vignette universe defined by factorial combinations of fixed levels of a finite number of dimensions. Experimental vignette studies also vary more widely in a presentation format; vignettes may be sampled or selected from an available pool of stimuli rather than constructed. This allows for additional creativity and flexibility

[1]It is worth noting that written SJTs are considered low-fidelity simulations in the personnel selection literature; however, this is relative to high-fidelity simulations such as assessment centres (Corstjens et al., 2017; Lievens et al., 2008).

in the design of a study – which may lead to more realistic stimuli and responses (i.e., greater fidelity). Repeating the presentation of some or all vignettes to different people provides the opportunity to compare responses to the same vignette between participants (Aguinis & Bradley, 2014; Finch, 1987).

Unfortunately, relaxing the constraints of factorial survey experiments changes the statistical model in important ways. First, it means that variability between vignettes is not properly accounted for by the factorial combination of vignette dimensions. Combinations of dimensions critical to estimating the effects of interest may be absent or under-represented. Second, responses to repeated presentations of the same vignette are not independent. This may be a relatively minor issue if vignettes *only* vary in terms of dimensions manipulated in a factorial design. This is because the shared variation in responses is fully accounted for the effect of individual dimensions or interactions between dimensions (provided they are included in the model). However, this is potentially a much more serious concern if there is variation in vignette features not associated with specific dimensions.

In practice, the distinction between experimental vignette designs and factorial survey designs is not always clear-cut. Thus, although experts generally recommend using a statistical model for nested repeated measures, this advice may not be sensible if the design does not meet the strictest definition of a factorial survey experiment (Auspurg & Hinz, 2015; Hox et al., 1991). For example, it is commonly advised to eliminate or reduce the number of implausible vignettes (creating imbalance). Furthermore, if the vignette universe is large, it will not be possible to allocate all combinations of vignette dimensions across participants without unreasonably large participant samples (Atzmüller & Steiner, 2010; Auspurg & Hinz, 2015).

The case for a general framework for the analysis of vignette data

The following sections consider the design and statistical modelling of vignette studies with a particular focus on minimizing bias and correctly accounting for variability in participants and vignettes. The discussion begins with consideration of common approaches to statistical modelling of vignette data, followed by consideration of how to handle incomplete sampling of the vignette universe and the distinction between nested and crossed factorial designs. This includes discussion of the potential hazard of unmodeled heterogeneity among vignettes. A potential solution in the form of a general framework for the analysis of vignette data via multilevel models is proposed. This is illustrated with an example of handling the complications of crossed random effects with rating data.

COMMON APPROACHES TO MODELLING VIGNETTE DATA

In a nested design, a different subset ('deck') of vignettes is presented to each participant (with no overlap between decks). If each participant makes only a single response to a single vignette, each observation is independent and data are straightforward to analyse. Independent measures ANOVA or ANCOVA is an obvious choice for such a factorial design (Aguinis & Bradley, 2014). Wallander (2009) reviewed 106 factorial survey experiments published in sociology journals and found that while the median number of vignettes per respondent was eight and the maximum 110, 21% of studies presented a single vignette. Although surprisingly common, single vignette studies are not generally desirable. They are straightforward to analyse but are extremely inefficient; they have low statistical power. Thus, many vignette studies adopt a nested repeated measures design in which each participant responds to more than one vignette, but a completely different deck of vignettes is presented to each participant. In such designs, the responses from the same individual are not independent – sometimes termed as intra-rater or intra-respondent correlation (Auspurg & Hinz, 2015; Rossi & Anderson, 1982; Wallander, 2009). Intra-respondent correlation is a potentially serious violation of the assumptions of the generalized

linear model – a family of models that include ANOVA, multiple linear regression and logistic regression (Baguley, 2012).

Wallander (2009) found that although intra-respondent correlation was not an issue for 28% of studies she surveyed, 33% (i.e., 46% of those where it was an issue) neither discussed nor addressed the problem, and only 21% (29% where it was an issue) adopted a statistical approach that accounted for it. The remaining studies adopted a variety of idiosyncratic approaches – with mixed likelihood that they fully addressed the problem. While there is no similar survey of analytic approaches for vignette data in psychology it seems likely that this problem is relatively widespread given that psychologists favour experiment vignette studies. This may be exacerbated by claims in the vignette literature that lack of independence is generally not a serious problem.

There are two aspects to this claim, which may sometimes be confused: correlation between responses from the same person and order effects. Order effects are relatively easy to deal with. For example, Rossi and Anderson (1982, p. 33) state 'serial order dependency among judgements may appear in some data sets, although […] in this volume serial order correlations within respondents appear not to be present to any appreciable extent'. Other research suggests strong order effects can arise but are not inevitable (Auspurg & Jäckle, 2017). Crucially, order effects can be controlled via counterbalancing or random allocation of vignettes to serial position. As counterbalancing is often impractical, randomization of presentation order is generally recommended (Auspurg & Hinz, 2015).

In contrast, if intra-respondent correlation is ignored this always results in an incorrect statistical model. The implications of ignoring the correlation between repeated observations from the same individual are well understood. Yet there are some puzzling claims in the literature about the viability of ignoring violations of independence. For example, (O'Toole et al., 1993, 1999) justify the use of ordinary least squares (i.e., a regression model that assumes independent, identically distributed errors) on the basis that the assumptions of a factorial survey design are met. However, there is nothing inherent in a factorial survey design that waives the independence assumption. The basis for this belief may be that correlations between vignette dimensions were generally close to zero and thus were near-orthogonal. While this is an important consideration that impacts the efficiency of the model, it does not address the problems arising from incorrectly treating correlated observations as independent. To understand why needs a consideration of the role of balance in a factorial design.

Orthogonality, intra-respondent correlation and imbalance

Orthogonality (independence of vignette dimensions) and intra-respondent correlation have distinct impacts on the appropriate choice of the statistical model. An important consideration in this context is whether the design is balanced. A balanced design is one where all dimensions are uncorrelated (orthogonal). The simplest way to ensure the balance is to have the complete vignette universe presented to each participant. If this is not feasible, balance can also be achieved relatively simply by presenting each possible vignette once but to different participants (e.g., if the vignette universe consists of 4096 vignettes one could randomly allocate eight different vignettes to each of 512 participants).

Repeated measures designs do not directly impact orthogonality but may make imbalance more likely and slightly more difficult to handle. This is because anything that disrupts the ideal allocation of vignettes to participants will cause departures from orthogonality and hence some collinearity between vignette dimensions. For instance, either dropping vignettes with implausible level combinations or missing responses will introduce imbalance. Randomly allocating vignettes to participants, a common characteristic of factorial survey designs with a large vignette universe (Auspurg & Hinz, 2015; Wallander, 2009), will also usually introduce imbalance. Unbalanced designs are undesirable because they introduce partial or complete confounding of effects. Where confounding arises it is not possible to uniquely attribute variation in the responses to the effect of a dimension or interaction between dimensions. This decreases efficiency (i.e., statistical power) as the estimates of some effects are

collinear under imbalance. It may also bias estimates relative to a completely balanced design (Auspurg & Hinz, 2015; Baguley, 2012; McCulloch, 2005).

While repeated measures designs do not inherently preclude orthogonality of effects, the most familiar approach for analysing such data, repeated-measures ANOVA, requires complete balance. This may increase the temptation, when confronted with data that aren't amenable to this analysis, to ignore the intra-respondent correlation and adopt approaches such as multiple linear regression that treat observations as independent. This is supported by claims in the literature that as long as the effects of interest are orthogonal this approach will produce identical regression estimates (Ludwick et al., 2004; Taylor, 2005; Taylor & Zeller, 2007). Some researchers also check for the impact of collinearity by confirming that the correlations between vignettes are low (O'Toole et al., 1993; St John & Heald-Moore, 1995) or by reanalyzing data with a balanced subset of vignettes (Rossi & Nock, 1982; St John & Heald-Moore, 1995). By and large, however, these checks are unnecessary. Provided the design is balanced the regression estimates themselves are not influenced by collinearity. The impact of collinearity is confined to reducing the effective sample size to assess the unique contribution of each predictor (Baguley, 2012; Goldberger, 1991).

Collinearity and effective sample size

This impact of collinearity is illustrated using measures such as the *VIF* (variance inflation factor) and tolerance. Tolerance is the proportion of variability in a predictor that is not shared with the other predictors in the model. It is therefore 1 if all predictors are orthogonal (they have exactly zero correlation) and 0 if a predictor can be perfectly predicted from all other predictors. This means that it indexes the effective sample size for that effect – for example, tolerance of 0.80 would indicate that the effective sample size is 80% of the nominal sample size.

As the standard errors (SEs) for inferences about an effect depend on effective sample size, this implies that they will be larger for an unbalanced design that introduces a degree of collinearity than for the equivalent balanced design. This can be assessed via the VIF statistic (where VIF = 1/tolerance). This indicates the ratio of the sampling variance (the square of the SE) in the unbalanced design relative to its balanced equivalent – for example, if tolerance = 0.80 then VIF = 1/.80 = 1.25 and thus the *SE* for that effect is 'inflated' by $\sqrt{\text{VIF}} = \sqrt{1.25} \approx 1.118$.

Thus collinearity introduces a cost in terms of efficiency (lower statistical power) even if collinearity is modest. Hence the clear guidance in the factorial survey literature to obtain a balanced or near-balanced design (Auspurg & Hinz, 2015). Fortunately, having the odd missing item or removing a few implausible items from a large vignette universe should not generally cause substantial imbalance (Auspurg & Hinz, 2015; Rossi & Nock, 1982; Wallander, 2009). However – as addressed below – the consequences for efficiency can be profound if the design is very unbalanced.

Repeated measures as special cases of clustered samples

The issue with repeated measures designs cannot, however, be reduced to whether vignette dimensions are orthogonal as this would completely ignore the impact of the intra-respondent correlation. In a typical study, our intention is a statistical generalization not only just to the vignette universe but also to the participants we are sampling. Ignoring the correlation among participants incorrectly treats each new observation as providing additional independent information to the model. This can distort estimates in an unbalanced design and (even in a balanced design) inflates the effective sample size, which, in turn, leads to underestimation of the SEs. In this sense, repeated-measures designs are a special case of the more general problem of clustered samples.

Kish (1965a, 1965b) provides a detailed explanation of how ignoring such clustering can lead to serious misinterpretations of findings. This is easiest to see by considering extreme cases. First, it

amounts to assuming that 1000 observations from 1 person, 100 observations from 10 people or 1 observation from 1000 different people are informationally equivalent. This is incorrect unless there is exactly zero variation between participants. Second, with unequal numbers of observations per participant, the estimates will predominately reflect those of large clusters. This could be desirable in some contexts but will be highly misleading in most situations (Kish, 1965a). Furthermore, it is nearly always undesirable for repeated measures designs where there is no reason to arbitrarily weight responses from some individuals more than others. Furthermore, it is now well-established that apparently modest degrees of clustering can have dramatic effects on the effective sample size. Hence ignoring clustering will lead to Type I error inflation (because SEs are systematically underestimated).

The impact on the effective sample size is encapsulated by what is termed a *design effect* (Kish, 1965b; Skinner, 1986). The degree to which error is inflated if clustering is ignored is given by the relationship $1 + (m - 1)\rho$ where ρ is the intraclass correlation coefficient (ICC) measuring the degree of clustering within units and m is the number of observations (here vignettes) nested within each unit (here participants).[2] When $\rho = 0$ there is no clustering, observations are independent and the design effect is 1. Although intraclass correlations can vary widely in value for different contexts, research from educational contexts suggests a range of 0.05 to 0.25 might be a reasonable starting point (Hedges & Hedberg, 2007, 2013). If anything, we might expect repeated measures designs (where clustering is within-person) to show greater degrees of clustering than children in the same class or school. For a hypothetical study with 500 participants and 12 vignettes per person and a modest ICC of $\rho = 0.05$, the design effect is 1.55. This translates to an effective sample size of $500 / 1.55 \approx 322.6$ and SEs that are too small by a factor of $\sqrt{1.55} \approx 1.25$. Even conservative estimates of the likely ICC in a repeated measures design would therefore have a material impact on the analysis of vignette data. Higher, more realistic, values of ρ and a larger number of vignettes would lead to more substantial impacts. It therefore seems likely that many published vignette studies have spuriously high statistical power; they underestimate the true variability in their data.

Incomplete or unbalanced sampling of the vignette universe

It is worth considering the role of balance in the design of a vignette study in further detail. In a purely statistical sense imbalance does not bias the estimates that are obtained – they remain unbiased estimates of the population (vignette universe) sampled. However, if the research questions of interest pertain to the balanced universe of vignettes then one can reasonably consider the estimates from an unbalanced design to be 'biased' estimates of the population of interest. This is analogous to sampling issues that arise in other contexts. For instance, a domestic violence researcher might survey individuals who have sought psychological support for domestic violence. This would be a suitable sample to make inferences about certain research questions (e.g., the characteristics and experiences of individuals who seek support) but is likely to lead to biased inferences if one is interested in the population of people who experience domestic violence (including those who did not seek support).

Imbalance can arise in a vignette study because some combinations of dimensions are not presented or because some combinations are oversampled relative to others. It may also be a structural feature of the way vignettes are allocated to participants. A concern in many vignette studies is that participants are asked about illogical combinations such as 'long job tenure' with 'no job experience' (Auspurg & Hinz, 2015) or implausible combinations such as low age and high monthly income. A common strategy in these cases is to drop such vignettes from the deck (Auspurg & Hinz, 2015; Wallander, 2009). For implausible combinations, it is less clear cut, though it may be desirable to reduce their number or to remove them. However, as

[2]For illustrative purpose, but without loss of generality, this example uses the common formula for a balanced design with two-stage sampling (Skinner, 1986).

the concern is that implausible vignettes may distort the findings it may be more sensible to account for it directly within the model (a consideration that will be explored in a subsequent section).

Wallander (2009) also suggests obtaining plausibility ratings from participants for each vignette as an option. This would allow researchers to restrict analyses to more plausible vignettes or (although not proposed by Wallender) to model the impact of plausibility on the estimates by including dimension × plausibility interaction terms.[3]

In many research contexts, while illogical or implausible combinations are not a major consideration, unbalanced designs nevertheless arise. This tends to occur when the vignette universe size exceeds the number of vignettes presented to participants. Under these conditions, the common strategy of randomly allocating vignettes to participants will in aggregate lead to near balance in reasonably large samples. This can be improved by allocating vignettes without replacement – ensuring that a greater proportion of the vignette universe is covered across all participants (and perfect balance if responses are obtained from every possible vignette an equal number of times). There are three main drawbacks of random allocation: (i) it may be too resource-intensive and therefore impractical to create bespoke combinations of vignettes for each participant (e.g., if using video), (ii) random allocation is not particularly efficient in terms of statistical power, and most critically (iii) partial or complete confounding may occur.

A solution for the first problem is to create decks of vignettes shared by multiple participants to ensure reasonable coverage of the vignette universe (Auspurg & Hinz, 2015). Efficiency can be increased by allocating vignettes to decks systematically in order to optimize statistical power to detect the effects of interest (Atzmüller & Steiner, 2010). This can also reduce partial and complete confounding, though it may not be possible to eliminate all confounding within the practical constraints of a specific study. We have already addressed the consequences of partial confounding, as this is merely the collinearity introduced by an unbalanced design.

Because vignette dimensions are correlated, their unique contribution to the estimation of the coefficients in the model is reduced (and hence tolerance <1 and VIF >1). Generally, this is not a huge problem for estimating the main effects of vignette dimensions, but for interactions between dimensions, statistical power will depend heavily on sampling-specific combinations of vignettes (McClelland & Judd, 1993). If these vignettes are entirely absent this results in complete confounding (though it is quite possible that there will be extremely low statistical power to detect some interaction effects even if complete confounding is avoided).

Complete confounding, fractional factorial designs and aliasing

Complete confounding occurs if one presents only a subset of the vignette universe to each participant the design, and this renders the design *fractional factorial* rather than *full factorial*. In a fractional factorial design, not all combinations of the dimensions (factors) are presented and some higher-order effects become *aliased* with lower-order effects (Auspurg & Hinz, 2015; Kirk, 1995). If two effects are aliased they cannot be distinguished from each other in the statistical model (and one cannot normally estimate both effects in the same model). This is not inevitable but can arise through resource or other constraints, a happenstance in random sampling (particularly if vignettes are allocated randomly to decks or the number of participants is small) or because of missing data. Fractional factorial designs have a long history in statistics and arise commonly in circumstances where there are physical limits on the number of observations per sampling unit (e.g., the number of machines or production lines in a factory).

To understand why aliasing is a problem, consider the example of a 3-way factorial ANOVA design. This is defined by a matrix in which each effect in the model is represented by a particular pattern of means termed a *contrast*. The common practice in regression is to represent the effects using values 0 and 1 (i.e., the contrasts use what is known as dummy coding). For a factorial design, however, contrast codes that sum to zero are often used, and a form of effect coding such as −1 and +1 is used. Table 1

[3]This requires a modelling approach, such as a multilevel modelling framework proposed here, that can incorporate time-varying covariates.

TABLE 1 An Example of the mapping between dimension (factor) levels and contrast codes for a 2×2×2 FACTORIAL DESIGN

Vignette	Levels of *A*	Levels of *B*	Levels of *C*	A	B	C	A×B	A×C	B×C	A×B×C
1	Male	Bullying	Public	−1	−1	−1	+1	+1	+1	−1
2	Male	Bullying	Private	−1	−1	+1	+1	−1	−1	+1
3†	Male	Harassment	Public	−1	+1	−1	−1	+1	−1	+1
4†	Male	Harassment	Private	−1	+1	+1	−1	−1	+1	−1
5†	Female	Bullying	Public	+1	−1	−1	−1	−1	+1	+1
6†	Female	Bullying	Private	+1	−1	+1	−1	+1	−1	−1
7	Female	Harassment	Public	+1	+1	−1	+1	−1	−1	−1
8	Female	Harassment	Private	+1	+1	+1	+1	+1	+1	+1

Note: Presenting a subset of vignettes for example † means that some effects may be aliased (e.g., here it would make it impossible to estimate both C and A x B x C if only these rows were included.

(adapted from (Auspurg & Hinz, 2015) sets out the contrast matrix for a $2 \times 2 \times 2$ factorial design that could represent a vignette design in three dimensions, each with only two levels (and thus a vignette universe of size of $2 \times 2 \times 2 = 8$).

Each vignette varies on three dimensions (A, B and C) representing, for example, the protagonist gender (male or female), a type of behaviour (bullying or harassment) and a location (private or public). Thus in the first vignette the contrast codes −1, −1 and +1 map into the levels 'male', 'bullying' and 'public' of dimensions A, B and C, respectively. The contrast codes for the 2-way and 3-way interaction effects ($A \times B$, $A \times C$, $B \times C$, $A \times B \times C$) are then simply the products of the respective codes for A, B and C (e.g., see Baguley, 2012). If all eight vignettes are presented (either to each participant or across participants) it is possible to estimate all effects including the 2-way and 3-way interactions. This full factorial design is orthogonal, and if the levels are balanced (each vignette and hence level appearing equally often), all effects are uncorrelated (i.e., collinearity is exactly zero). Note there is still an advantage of presenting as many vignettes as reasonably possible to each participant because this produces more data, but provided there is complete balance and there is no loss of efficiency from the way the vignettes are allocated. Indeed allocating systematically to ensure balance is far more efficient than random allocation (Auspurg & Hinz, 2015).

Serious problems can arise if the design is fractional rather than full factorial. For instance, what happens if we only presented vignettes 3, 4, 5 and 6 rather than all 8 vignettes from the example in Table 1? Levels of dimensions A, B and C are still balanced and therefore we can still estimate the main effects of gender, behaviour and location, but we cannot estimate some effects at all (e.g., $A \times B$) and the main effect of C is completely aliased with the 3-way interaction $A \times B \times C$. To see why, one can compute the correlation coefficient between the contrast codes for the relevant contrasts. For the full set of 8 vignettes the correlation between the codes for C and $A \times B \times C$ is 0, but for the fraction including only vignettes 3 to 6, the correlation is −1 (a perfect negative correlation).[4] Within that fraction we cannot estimate the effect of C without it being aliased (completely confounded) with the 3-way interaction (and we cannot include both effects in the model). The precise pattern of estimable effects and aliasing depends on the design and the subset of vignettes in the fraction presented to participants.

[4]For example using R (R Core Team, 2020) cor(c(1,−1,1,−1,1,−1,1,−1), c(−1,1,1,−1,1,−1,−1,1)) will return 0 and cor(c(−1,1,−1,1), c(1,−1,1,−1)) returns −1.

Dealing with aliasing of effects

A common strategy to deal with aliasing is to estimate only the main effects and only some (lower order) interaction effects. For instance, for the example in Table 1 one could estimate just the main effects with vignettes 3 to 6. If the interactions between dimensions were negligible these estimates would be largely unbiased estimates of the parameters in the full factorial design. Higher-order interaction effects typically do not tend to account for a large proportion of variation and ignoring these effects is frequently defensible.[5] However, two-way interactions are often of particular theoretical interest in psychology research and being unable to uniquely identify their effects is probably highly undesirable. For instance, it might be particularly interesting to know whether the impact of bullying and harassment differs between genders or locations (e.g., it would be rather surprising if their effects were the same for male, female and nonbinary individuals).

Given that it is not always possible or realistic to have a full factorial design it is crucial to appreciate that there are ways to optimize the efficiency of fractional designs and to limit or avoid aliasing of effects. The approach recommended by Atzmüller and Steiner (2010) and Auspurg and Hinz (2015) is to select vignettes based on their *D-efficiency*. D-efficiency is optimized by maximizing the determinant of the covariance matrix and hence minimizing the error in the model. In this approach, rather than selecting a vignette fraction that aliases only higher-order effects (assumed to be ignorable), one tries to find a fraction that is near optimal in terms of D-efficiency; the efficiency relative to the desired model (e.g., a main effects only model or one with main effects plus two-way and three-way interactions). Thus D-efficiency of 1 is optimal, whereas D-efficiency of 0.72 is 72% as efficient as the model of interest. If the vignette universe is large (which is likely for applications where high D-efficiency is desirable) then determining an optimal or near-optimal allocation of vignettes is laborious. Researchers therefore rely either on known designs or use computer software to search for combinations of vignettes that maximize D-efficiency under various constraints. Although full treatment of D-efficiency is beyond the scope of this paper, a simple example of how to obtain a D-efficient allocation of vignettes is provided in Appendix A.

THE FIXED-EFFECT FALLACY AND NESTED VERSUS CROSSED DESIGNS

So far the discussion has focused primarily on nested designs. In its simplest form, a nested design involves one random factor corresponding to the sampling unit. For psychology research, this is typically people (but might be animals, teams etc. depending on context). Treating a sample of participants in a study as a random effect in a statistical model allows a researcher to make inferences about the population from which that sample is drawn. Observations in such studies can also vary as a function of the properties of stimuli (items) that are presented. These are often represented as predictors (factors or covariates) included as fixed effects in the model. The core distinction between a random effect and a fixed effect is that a fixed effect exhausts (or at least substantially depletes) the population of interest while a random effect does not. For example, participants are typically a random effect because we are interested in generalizing to a potentially infinite population of similar people, while a variable such as marital status is typically fixed because we wish to compare a finite number of categories.

Vignette studies present an intriguing challenge in this regard because it is not always appropriate to treat them as nested designs. Specifically, it is only appropriate to treat a vignette as a nested design if: (i) each participant receives a unique set of vignettes (and thus vignettes are also nested), or (ii) the

[5]This is sometimes known as the hierarchical ordering principle. While not inevitable, it is a well-known regularity in empirical research deriving both from the structure of the world and the way that researchers design their experiments—see Li et al. (2006).

dimensions that define the vignettes exhaust the variability in the vignette universe. If these conditions do not hold then the vignette study will fall victim to the fixed-effect fallacy.

The fixed effect fallacy was identified by Coleman (1964) in the context of psycholinguistic research and later popularized by Clark (1973). At that time most psycholinguistic research treated participants as random effects but samples of language stimuli as fixed effects. This has potentially catastrophic consequences for statistical modelling because the variability between items (typically words in psycholinguistic research) is ignored when analysing means of items within the same experimental condition. Underestimating or ignoring variability between items leads to liberal inferences. For example, SEs will be underestimated, and hence, Type I error inflated. With nominal

$$\alpha = 0.05$$

the true Type I error might easily exceed 0.60 or more if items are incorrectly treated as fixed factors, under fairly common conditions (Judd et al., 2012).

How likely is this to be a problem in most vignette studies? Factorial survey designs in sociology have from the beginning recognized the importance of capturing the full variability of the vignette universe (Rossi & Nock, 1982). In principle, the fixed-effect fallacy should not therefore apply to the majority of factorial survey designs. However, in practice even relatively strict factorial survey designs incorporate additional variation in the surface form of the vignettes presented. This can be seen in the sample vignette from (Rossi & Anderson, 1982) presented earlier, where the names of the characters ('Cindy M.' and 'Gary T.') vary between vignettes. Indeed, Wallander (2009) explicitly recommends adding several additional dimensions that are not considered predictors in the model as a way to make vignettes seem less repetitive and more plausible, citing a study by Shively (2001) as a blueprint. Evans et al. (2015, p. 162) term these 'contextual aspects' that can be added to vignettes 'in order to provide verisimilitude (e.g., nonessential details that enhance the 'personhood' of a vignette character), but are not thought to exert a causal influence on the dependent variables.' It thus seems likely that this kind of *unmodeled heterogeneity* can arise even in a carefully designed study. Furthermore, for many sources of unmodeled heterogeneity–such as the names in the sample vignette above or those involving nontext presentation formats such as images or videos–the size and variability of the true vignette universe will be undefined or unknown.

DEALING WITH UNMODELED BETWEEN-VIGNETTE HETEROGENEITY

Unmodeled between-vignette heterogeneity is challenging to the analysis of vignette data. Any model that does not handle this variability appropriately will underestimate sampling error (with consequences such as Type I error inflation noted earlier). In addition, accounting for the heterogeneity by altering the design may not be practical because the factors causing vignettes to vary are not sufficiently well understood. Fortunately, the problem has been widely studied in the context of the fixed-effect fallacy and four broad approaches to handling the problem can be identified:

1. Acknowledge the limitations on the generalizability of the results. If there is unmodeled between-vignette variation that cannot be ignored, then any inferences from statistical modelling are restricted to the sample of vignettes that were presented.
2. Propose that between-vignette heterogeneity can be ignored. This could be argued if the relevant features are modelled near-exhaustively via the inclusion of fixed effects, on the basis of previous research that suggests such variation is negligible, or that the features underlying the variation are a superficial 'contextual aspect' (Evans et al., 2015).
3. Sample vignettes at random from the full universe of interest so that each participant receives a different random selection of items (e.g., see Clark, 1973). This is only possible if the full universe

is available for sampling. It also has important drawbacks. It may limit the factors that can easily be manipulated in the vignette study (e.g. with video stimuli) and relying on the random sampling of vignettes tends to produce less efficient designs (Auspurg & Hinz, 2015). It may also be much more resource-intensive to set up the study (though this can be minimized with common experiment presentation software).

4. Incorporate between-vignette heterogeneity in the statistical model. An elegant way to handle this is to incorporate vignette as a random effect in a multilevel model[6] (Baguley, 2012; Judd et al., 2012). This is a very flexible analytical framework that can handle many different types of response and incorporate random slopes to model individual differences in the size of an effect. It has a number of attractive features for analysing vignette data. Notably, it can cope with fully crossed designs where each participant receives the same vignettes or partially crossed designs where only some vignettes are shared between participants. It can also handle incomplete and fractional factorial designs. A final feature of the multilevel approach is that estimates of the random effects of predictors are partially pooled across participants and across vignettes. They are therefore shrinkage estimators (Baguley, 2012; Greenland, 2000) in which the vignette and participant estimates are shifted closer to their average. Thus when we have more information about some items or individuals than others (e.g., because of imbalance) the resulting estimates–including fixed effects–borrow strength from each other.[7]

A full discussion of shrinkage is beyond the scope of this paper, but in the context of a vignette study, shrinkage is particularly attractive as it will reduce the impact of implausible vignettes (and atypical participants). Other approaches might also be adopted – notably generalized estimating equations (GEEs) or cluster-corrected SEs (McNeish et al., 2017). However, these forms of analysis are not as flexible as the multilevel approach (e.g., in handling crossed random effects) and do not incorporate shrinkage.

Deciding on which analytic approach to adopt

For most research the first approach is unlikely to be satisfactory; the reason for adopting a quantitative vignette study is to generalize beyond the vignettes presented and to make inferences about broader behaviour. The second approach is more reasonable but can be hard to defend without specific empirical support.

One should, in particular, be careful about claiming that the features underlying the variation in vignettes are superficial. Consider the case of the names assigned to characters in the sample vignette presented earlier. At one level these are entirely superficial, so how likely is that they could influence responses to the vignettes? Although there is a considerable literature in cognitive psychology that suggests proper names lack (or at least are not processed for) meaning this is at odds with the broader societal, cultural and personal significance that names play in everyday life (Brennen, 2000; Pilcher, 2016). Brennen (2000) resolves by arguing that names vary in meaningfulness between people and contexts but are less likely to be processed for meaning with repeated exposure. It is possible to extend this logic to the context of a vignette study in which the most salient factors likely to influence responses are strictly controlled (by including them as dimensions or by standardizing them). Under these constraints superficial labels and features, especially when encountered for the first time, are likely to be more influential than they would be in richer everyday life contexts.[8] For instance, in our earlier example we know the age of 'Gary T.' but not 'Cindy M.' (and the ethnicity of neither). Absent this kind of information, participants might react differently to the

[6]Multilevel models are also known as linear mixed models or hierarchical linear models in the literature.

[7]This way of thinking about shrunken estimates as 'borrowing strength' is derived from John Tukey's early work in this area (Brillinger, 2002).

[8]This is similar to the minimal group paradigm in social psychology, which show large effects of superficial labels precisely because other group identity information has been stripped out (e.g. see Reicher, 2004).

vignette if the protagonist were named 'Doris J.' or 'Rana A.', which suggest very different demographic profiles.

The choice between the remaining approaches will therefore depend on trade-offs between competing priorities. Randomly allocating different samples from the vignette universe to each participant is likely to be more resource-intensive than including vignette as a random effect in a multilevel model but may lead to greater generalizability. Relative efficiency is harder to assess as it depends on the total number of unique vignettes, and the impact of randomly sampled vignette dimensions on D-efficiency. However, it is possible to combine the best of both approaches by designing or selecting a large number of vignettes that are representative of the full vignette universe. For example, one could tightly control the key dimensions and randomly sample names of characters from a diverse database. Each vignette is therefore made up of a basic template or recipe plus randomly sampled features and vignette dimensions that are manipulated systematically. At this point, whether there is unmodeled heterogeneity becomes an empirical question. By fitting a multilevel model with random intercepts for the participants and for the vignettes, it becomes relatively trivial to assess whether there would be unmodeled heterogeneity from either source in a simpler model. At that point, it is possible but not necessary to switch to the simpler model as the two models should provide identical or near-identical inferences.

The rationale for this approach is essentially the same as for selecting a fixed or random effects model in meta-analysis. The fixed-effects model in meta-analysis assumes that each study is estimating a constant fixed population effect size. If the effect size is not constant but varies between studies, there is unmodeled heterogeneity. This can be assessed in a random-effects model, though in many contexts a fixed-effects model is considered implausible a priori. Just as moderators in a meta-analysis can reduce heterogeneity, adding fixed factors and interactions for the vignette dimensions or participant characteristics may reduce heterogeneity associated with the vignettes or participants. Best practice in both cases is to start by fitting a random-effects model to explicitly model variation in the units of interest (studies or vignettes and participants). In this sense, there is no cost to adopting the random-effects model (see Borenstein et al., 2010). An example of this approach is provided in the next section, with the corresponding R code in Appendix B.

EXAMPLE: MULTILEVEL ANALYSIS OF RATING DATA WITH CROSSED RANDOM EFFECTS

Multilevel models provide a general and flexible approach to handling a wide range of quantitative vignette data. An important consideration, however, is the form of the vignette response. This may be in the form of a dichotomous outcome or a visual-analog scale but is most commonly an ordinal rank or rating. While it is common to treat ordinal responses as interval, there is now increasing awareness that this can lead to serious problems (Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018). Fortunately, dichotomous data can be modelled using multilevel logistic regression and rating data using multilevel ordered logistic regression.[9] The following example therefore demonstrates how to analyse vignette rating data with random intercepts for participants and items (vignettes) using the ordinal package in R (Christensen, 2019). It takes a frequentist approach to statistical inference, but Supplementary Materials the online resources accompanying this article (https://osf.io/q2g63/) include a re-analysis in R using brms (Bürkner et al., 2020; Bürkner & Vuorre, 2019) which allows for a wider range of models and responses, and the ability to incorporate informative priors and obtain Bayes factors and posterior probability (credibility) intervals if desired.

The context of the example is a study comparing the perceived severity, and willingness to report bullying and harassment behaviours online or offline with a student sample (Dunham, 2018). The vignettes were created using a stem and leaf format adapted from (Sticca & Perren, 2013). Each stem described an instance of bullying or harassment (e.g., 'A neighbour in your halls of residence reads insults about their home life') accompanied by a leaf providing a different offline or online context (e.g., 'via posters

[9]For visual-analog scales the best approach depends somewhat on the pattern of responses and, in particular, whether there are 'excess' responses at the extremes of the scale. If not, a normal model may be adequate. Otherwise either treating the responses as a censored normal distribution or a mixture model such as a zero–one inflated beta-binomial would be sensible options.

TABLE 2 Ordinal logistic regression coefficients for an intercept-only model with random effects of item and vignette and severity as an outcome

Random effects		Variance	*SD*
Participant		2.40	1.55
Vignette		0.94	0.97
Threshold	**Coefficient**	**SE**	**95% CI**
1/2	−5.10	0.32	−5.72, −4.47
2/3	−3.17	0.28	−3.72, −2.62
3/4	−1.06	0.27	−1.59, −0.53
4/5	1.24	0.27	0.71, 1.77

Note: Log-likelihood = −2174.73, Likelihood χ^2 = 4349.5.

around the university campus'). Participants read each of the 24 vignettes and rated them from 1 to 5 for the severity of and willingness to report the described behaviour. The basic analysis is reported here with the corresponding R code provided in Appendix S1 (and examples extending Appendix S1 in the supplementary material accompanying this article). This analysis modelled the severity rating with three fixed factors representing the dimensions Gender (male or female), Behaviour (bullying or harassment), Medium (offline or online) and two random factors (participant and vignette). Given that most vignette studies involve a factorial design our intention here is to mimic the structure of an ANOVA analysis.

If the outcome had only two values (0 and 1) this would be logistic regression in which the log odds of responding 1 are a linear function of the predictors. Conceptually one can think of an ordinal logistic regression as an extension to this model in which the log odds are modelled at each threshold between ratings (e.g., 1 vs. 2+, 1–2 vs. 3+ and so on). In this way, the model provides the cumulative log odds at each threshold. These thresholds are flexible by default and, unlike a model that treats the outcome as continuous, differences between ratings aren't forced to be equal in magnitude.

The initial step is to fit an intercept-only model–a model with no predictors but with random intercepts for both participants and vignettes. As the response is a 1 to 5 rating this model will have four intercepts (representing the cumulative logs odds of the threshold between successive ratings). The output of this model also gives the overall fit of the model (in terms of the likelihood chi-square) and estimates of the random intercept standard deviations (summarized in Table 2). The random effects indicate there is variation in the average ratings between participants and between vignettes, though participants account for the majority (72%) of the random effect variance. Although these estimates may increase or decrease if predictors are added to the model, we are particularly interested in whether the vignette variance (the unmodeled heterogeneity in vignettes) decreases when factors representing vignette dimensions are added. Table 2 also shows the cumulative log odds of each threshold increasing from −5.10 for the 1|2 threshold to 1.24 at the 4|5 threshold. The former represents a probability of $e^{-5.10}/(1+e^{-5.10}) = 0.0061e^{\neq 5.10}$ ¥ 0.0061 and the latter a probability of

$$e^{1.24}/(1+e^{1.24}) = 0.7756$$

e$^{1.24}$ ¥ 0.7756.

The next step is to fit a model with the main effects of interest followed by a model with all two-way interactions. To aid the interpretation of interaction models (and make model fitting more efficient), it can be helpful to use a form of effect coding (such as −1,1) for the factors (Baguley, 2012; Schad et al., 2020), but for present purposes, we adopt the more familiar dummy coding approach. We could also fit the three-way interaction model, but for many vignette studies, it is necessary or desirable to limit

TABLE 3 Likelihood ratio tests of main effects and two-way interactions with severity as the outcome

	df	LRT	*p*
Behaviour	1	9.61	.0019
Medium	1	0.03	.8708
Gender	1	0.96	.3281
Behaviour × medium	1	2.91	.0880
Behaviour × gender	1	6.81	.0091
Medium × gender	1	7.21	.0073

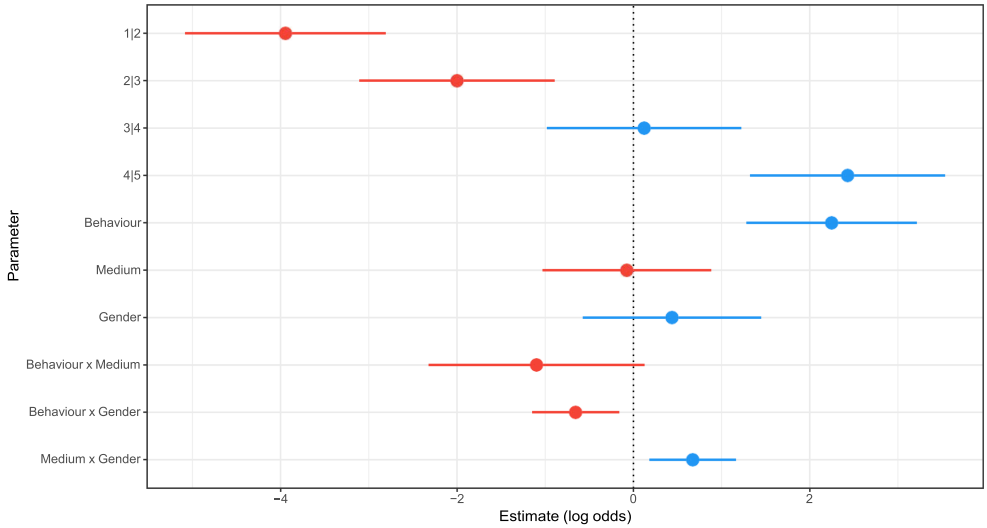

FIGURE 1 Fixed-effects parameter estimates with 95% CIs for the two-way interaction model with severity as an outcome. *Note*. All estimates are for categorical factors with dummy coding (0,1)

the model to lower-order effects (and here case the three-way model offers a negligible improvement in fit over the two-way model).[10] To obtain ANOVA-like tests of the main effects one can drop each main effect in turn from the main effects only model. Unlike ANOVA, this approach gives a likelihood ratio test (*LRT*) rather than an *F* test.

Likewise, one obtains tests of the two-way interaction effects by dropping each interaction in turn from the two-way model. This approach to testing effects is known as Type II (hierarchical) sums of squares and, in a balanced design, equivalent to the widely used Type III (unique) sums of squares approach implemented in SPSS and SAS, though the latter approach has attracted some criticism (e.g., see Nelder & Lane, 1995).

The tests of main effects and two-way interactions are summarized in Table 3. In this model, the *SD* of the random effect of the participant is unchanged (at 1.55) even though Gender and its interactions with the other factors were added to the model. The *SD* of the vignette random effect has, however, decreased to 0.73. This indicates although the behaviour and medium dimensions are accounting for some variability between vignettes, there is still unmodeled heterogeneity (and although not reported here, this remains when the three-way interaction is added). Thus, even with this relatively simple stem and leaf vignette format, it would be unreasonable to treat vignette as a fixed effect.

[10]It may also be necessary to limit the specific lower-order interactions if there are aliased effects.

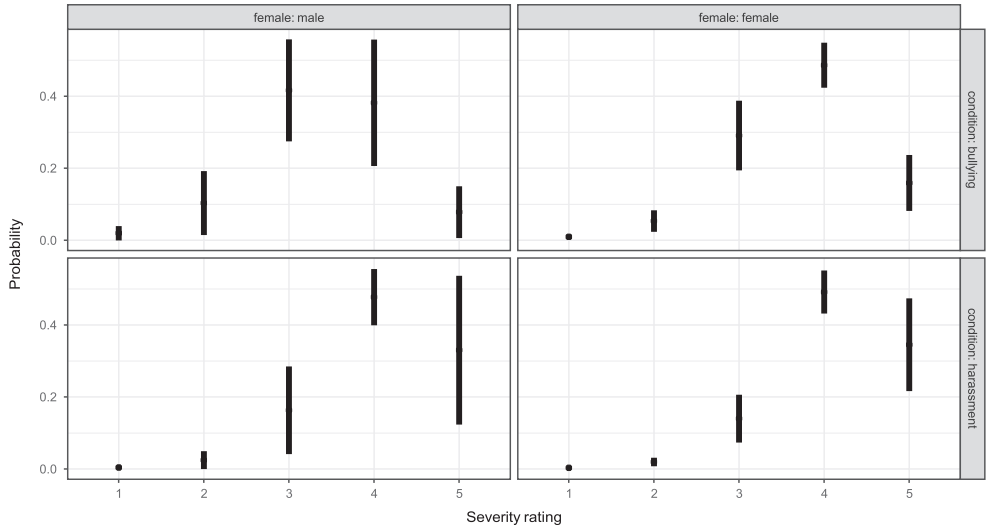

FIGURE 2 Predicted probability of severity ratings for the behaviour by gender interaction. *Note*. Error bars are 95% CIs

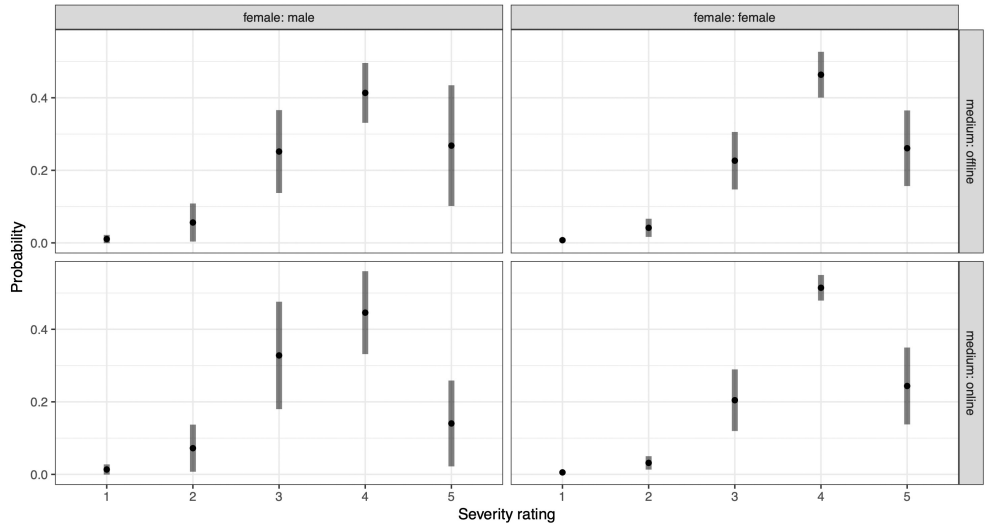

FIGURE 3 Predicted probability of severity ratings for the medium by gender interaction. *Note*. Error bars are 95% CIs

Figure 1 shows a coefficient plot of the two-way model with effect-coded factors. Interpretation of these effects on the log-odds scale is not straightforward, not least because there are statistically significant two-way interactions of Behaviour × Gender and Medium × Gender. In these cases interpreting the mean rating can be misleading because the probability of a higher rating depends on the other effects in the model (even if there is no interaction on the log-odds scale). Best practice is therefore to look at the shift in predicted probabilities averaging over other effects in the model. This also has the advantage of proving a richer description of the participant responses than just looking at the mean rating (which could conceal interesting patterns). These are shown in Figures 2 and 3.

Figure 2 reveals female students are more likely to rate vignettes describing bullying as more severe than male students, while both male and female students tend to rate harassment similarly. The plot also shows harassment tends to elicit more severe ratings than bullying (consistent with the main effect of behaviour type). For the medium by gender interaction, a more subtle pattern arises–males assign lower severity ratings than females, but only online. Male ratings are generally similar to those of females for offline but are considerably less likely to give the highest rating to online behaviour.

Extensions to multiple item responses and more complex models

The Supplementary Materials online resources accompanying this article (https://osf.io/q2g63/) extend the analysis presented to illustrate incorporating a second dependent variable, random slopes and using brms to obtain Bayesian inferences or deal with convergence issues in more complex models. Including random slopes may be of particular theoretical interest as it would allow a researcher to see if the effect of a dimension varied between participants or between vignettes. Multiple outcome variables present a further challenge for vignette studies where the approach often lends itself to collecting multiple responses per vignette. As these are likely positively correlated it can be difficult to develop a coherent analytic approach for handling multiple outcomes. The simplest approach is probably to analyse each outcome separately and incorporate a multiple comparison correction (e.g., Baguley, 2012; Hochberg, 1988) that takes into account the correlation. In some cases, it may be preferable to use a data reduction technique such as factor analysis to pool items into a single outcome score. However, the multilevel framework proposed here also allows multiple outcomes to be incorporated within a single model. This is illustrated Supplementary Materials online resources accompanying this article (https://osf.io/q2g63/) by extending our example to include both the severity and willingness to report ratings.

CONCLUSION

Vignette methods have the potential to provide rich data on social and psychological phenomena that are hard to study directly. Quantitative vignette studies nearly always aim to generalize beyond the vignettes a particular individual is presented with. To do so requires researchers to align their chosen statistical model to the structure of their data. Understanding the importance of the design of the vignette study to the selection of an appropriate statistical model will also help avoid unnecessary aliasing of effects if a fractional design is used or to allow more efficient allocation of vignettes to participants. Except in the unusual case that each participant is only presented with one vignette it will be necessary for the model to account for the dependency between observations nested within participants. If the vignette universe is sufficiently large it will be possible (though not always practical) to present different vignettes to each participant, leading to a nested design that can be handled with a model treating participants as a random effect. However, it is common in psychological research to present some or all vignettes to more than one participant. Under such circumstances, it is appropriate to treat the variability between vignettes as a random effect in the model. Having a source of unmodeled heterogeneity will lead to underestimation of the total error and hence problems such as Type I error inflation.

Multilevel models provide a suitable general framework to analyse data from nested and partially or fully crossed random factors. Other approaches are available, but multilevel models have three main advantages in this context. First, not all the alternative approaches can easily incorporate both crossed and nested designs (e.g., this is difficult with GEEs or cluster-corrected SEs). Second, the shrinkage of estimates in the multilevel towards that of a typical unit is highly desirable in vignette studies and will reduce the impact of implausible vignettes on the estimates. Third, quantitative vignette studies often have outcomes in the form of rating scales, discrete responses or visual-analog scales. Generalized multilevel linear models have the flexibility to model a wide range of response formats and to capture

other unusual properties of a vignette study (e.g., additional clustering within different organizations or multiple outcome measures).

How common are the issues raised here with respect to vignette studies? In theory, they should be relatively rare in the factorial survey experiment literature where effort is made to allocate vignettes without repetition across participants and the vignette sampling fraction approaches 100%. However, even in these studies, it seems common not to treat participants as a nested random effect (O'Toole et al., 1993, 1999). Furthermore, it is clear even then the vignette universe may not be exhausted by the dimensions being modelled (as in the example of adding names such as 'Cindy M.' and 'Gary T.'). These apparently irrelevant details serve to make the vignettes less formulaic and repetitive (Wallander, 2009), but at the cost of introducing unmodeled heterogeneity. Crossed designs also appear to be more common than nested designs in several disciplines (Aguinis & Bradley, 2014; Atzmüller & Steiner, 2010; Wason et al., 2002). Given there are other reasons to consider using a multilevel model for vignette data, it seems appropriate to treat the presence of unmodeled heterogeneity as an empirical question. If the between-vignette variability is truly negligible then the estimates from the model with crossed random effect will be equivalent to a model in which there is no random effect.

A further consideration is that issues with the design and analysis of vignette studies arise also in other contexts. Indeed, any study that presents participants with audio clips, video clips or VR scenarios (whether manipulated or not) and obtains responses from them could be argued to be a type of experimental vignette study. For example, hazard perception tests (Horswill, 2016) present multiple video clips or simulated driving scenarios in which participants have to respond to a hazard or potential hazard. However, only a relatively few recent studies treat stimuli as a random effect (e.g., Crundall et al., 2021; Ventsislavova et al., 2019). As it becomes cheaper and easier to manipulate video and computer-generated video stimuli it seems likely that complex study designs closer to text-based factorial survey experiments will become more common. One advantage of treating these experiments as vignette studies is to consider the potential for fractional designs and efficient allocation of stimuli to increase statistical power or the number of factors.

An important consideration in the generation of vignettes is the fidelity of vignette content to the real-world context a researcher is generalizing to. A statistical model can only go so far in ensuring generalizability–the vignettes (and participants) have to be representative of that context. The ideal here is either a random or stratified random selection (such as from a D-efficient design) from the population. This is not always possible. One could not easily sample people's bullying experiences or driving hazards at random. However, it may be possible to obtain reasonably large samples of such incidents and select or (to protect people's privacy) create vignettes that preserve the features of interest. Alternatively one could draw on domain experts to validate the vignette content (Lanza & Carifio, 1992). There is also a rich body of work on increasing the fidelity of scenarios in the SJT literature that can inform vignette design (Corstjens et al., 2017; Lievens, 2017). An important consideration here is that fidelity in the vignette content will often go hand-in-hand with heterogeneity in the vignettes. Thus eliminating between-vignette heterogeneity will probably not be desirable from a generalizability perspective, even though it will make statistical generalization more challenging (by adding to the overall error in the model).

Whether a crossed or nested design is adopted also has implications for the number of vignettes to present. Estimating statistical power and sample size for a multilevel model is not trivial (see Judd et al., 2017). However, just understanding it is a crossed rather than nested design is instructive. The participants and vignette effects are independent random effects that together contribute to the total error variance in the model. Thus if participants and vignettes are equally variable you would want as many different vignettes as participants in a fully crossed design. If you have used similar stimuli before you may have an idea of the relative variability of participants and vignettes that can inform your judgement (and ideally a simulation of statistical power). Increasing only the total number of participants or the total number of vignettes will provide diminishing returns because the total error in the model will be dominated by the smaller of the two sample sizes. For these reasons the efficiency of the design to test the most theoretically or practically important effects should not be neglected.

AUTHOR CONTRIBUTIONS

Thom Baguley: Conceptualization; formal analysis; investigation; methodology; resources; software; supervision; visualization; writing – original draft; writing – review and editing. **Grace Eve Nell Dunham:** Conceptualization; data curation; resources; writing – review and editing. **Oonagh Steer:** Conceptualization; investigation; methodology; resources; writing – review and editing.

CONFLICT OF INTEREST

All authors declare no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://osf.io/q2g63/.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the Supplementary Materials of this article. Supplementary Materials (including R code) are available via OSF. Thank you to Sam Baguley, Jens Roeser, Sarah Seymour-Smith and Clifford Stevenson for helpful input during the preparation of the manuscript.

ORCID

Thom Baguley https://orcid.org/0000-0002-0477-2492
Grace Dunham https://orcid.org/0000-0001-6476-0437
Oonagh Steer https://orcid.org/0000-0002-0922-5498

REFERENCES

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, *17*(4), 351–371.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology*, *6*(3), 128–138.

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. Sage.

Auspurg, K., & Jäckle, A. (2017). First equals most important? Order effects in vignette-based measurement. *Sociological Methods & Research*, *46*(3), 490–539.

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioural sciences*. Palgrave Macmillan.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111.

Brennen, T. (2000). On the meaning of personal names: A view from cognitive psychology. *Names*, *48*(2), 139–146.

Brillinger, D. R. (2002). John W. Tukey: His life and professional contributions. *The Annals of Statistics*, *30*(6), 1535–1575.

Bürkner, P.-C., Gabry, J., & Weber, S. (2020). Brms: Bayesian regression models using 'Stan'. Retrieved January 13, 2021, from https://CRAN.R-project.org/package=brms

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgement tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, *82*(1), 143–159.

Christensen, R. H. B. (2019). Ordinal: Regression models for ordinal data. Retrieved March 12, 2021, from https://CRAN.R-project.org/package=ordinal

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behaviour*, *12*(4), 335–359.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports*, *14*(1), 219–226.

Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational judgement tests for selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention* (pp. 226–246). Wiley.

Crundall, D., van Loon, E., Baguley, T., & Kroll, V. (2021). A novel driving assessment combining hazard perception, hazard prediction and theory questions. *Accident Analysis & Prevention*, *149*, 105847.

de Groot, T., Jacquet, W., De Backer, F., Peters, R., & Meurs, P. (2020). Using visual vignettes to explore sensitive topics: A research note on exploring attitudes towards people with albinism in Tanzania. *International Journal of Social Research Methodology*, *23*(6), 749–755.

Dunham, G. E. N. (2018). *Examining Perceived Severity of and Willingness to Report Bullying and Harassment Online and Offline* (Unpublished BSc Psychology Project). Nottingham Trent University Nottingham.

Evans, S. C., Roberts, M. C., Keeley, J. W., Blossom, J. B., Amaro, C. M., Garcia, A. M., Stough, C. O., Canter, K. S., Robles, R., & Reed, G. M. (2015). Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in ICD-11 field studies. *International Journal of Clinical and Health Psychology*, *15*(2), 160–170.

Finch, J. (1987). The vignette technique in survey research. *Sociology*, *21*(1), 105–114.

Goldberger, A. S. (1991). *A course in econometrics*. Harvard University Press.

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology*, *29*(1), 158–167.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87.

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*(6), 445–489.

Hine, B. (2019). 'It cannot be that bad, I mean, he's a guy': Exploring judgements towards domestic abuse scenarios varied by perpetrator and victim gender, and abuse type. In E. A. Bates & J. Taylor (Eds.), *Intimate partner violence: New perspectives in research and practice* (pp. 43–57). Routledge.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802.

Horswill, M. S. (2016). Hazard perception in driving. *Current Directions in Psychological Science*, *25*(6), 425–430.

Hox, J. J., Kreft, I. G. G., & Hermkens, P. L. J. (1991). The analysis of factorial surveys. *Sociological Methods & Research*, *19*(4), 493–510.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*(1), 601–625.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioural sciences* (3rd ed.rd ed.)). Brooks/Cole.

Kish, L. (1965a). Sampling organizations and groups of unequal sizes. *American Sociological Review*, *30*(4), 564.

Kish, L. (1965b). *Survey sampling*. Wiley.

Lanza, M. L., & Carifio, J. (1992). Use of a panel of experts to establish validity for patient assault vignettes. *Evaluation Review*, *16*(1), 82–92.

Lewis, G., Croft-Jeffreys, C., & David, A. (1990). Are British psychiatrists racist? *British Journal of Psychiatry*, *157*(3), 410–415.

Li, X., Sudarsanam, N., & Frey, D. D. (2006). Regularities in data from factorial experiments. *Complexity*, *11*(5), 32–45.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. *Journal of Experimental Social Psychology*, *79*, 328–348.

Lievens, F. (2017). Assessing personality–situation interplay in personnel selection: Towards more integration into personality research. *European Journal of Personality*, *31*(5), 424–440.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgement tests: A review of recent research. *Personnel Review*, *37*(4), 426–441.

Ludwick, R., Wright, M. E., Zeller, R. A., Dowding, D. W., Lauder, W., & Winchell, J. (2004). An improved methodology for advancing nursing research: Factorial surveys. *Advances in Nursing Science*, *27*(3), 224–238.

Maieron, M. J., Roberts, M. C., & Prentice-Dunn, S. (1996). Children's perceptions of peers with AIDS: Assessing the impact of contagion information, perceived similarity, and illness conceptualization. *Journal of Paediatric Psychology*, *21*(3), 321–333.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*(2), 376–390.

McCulloch, C. E. (2005). Repeated measures ANOVA, R.I.P.? *CHANCE*, *18*(3), 29–33.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modelling. *Psychological Methods*, *22*(1), 114–140.

Nelder, J. A., & Lane, P. W. (1995). The computer analysis of factorial experiments: In memoriam-frank yates. *The American Statistician*, *49*(4), 382.

O'Toole, A. W., O'Toole, R., Webster, S., & Lucal, B. (1993). Nurses' recognition and reporting of child abuse: A factorial survey. *Deviant Behaviour*, *14*(4), 341–363.

O'Toole, R., Webster, S. W., O'Toole, A. W., & Lucal, B. (1999). Teachers' recognition and reporting of child abuse: A factorial survey. *Child Abuse & Neglect*, *23*(11), 1083–1101.

Pilcher, J. (2016). Names, bodies and identities. *Sociology*, *50*(4), 764–779.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reicher, S. (2004). The context of social identity: Domination, resistance, and change. *Political Psychology*, *25*(6), 921–945.

Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgements: The factorial survey approach* (pp. 15–67). Sage Publications.

Rossi, P. H., & Nock, S. L. (1982). *Measuring social judgements: The factorial survey approach*. Sage Publications.
Sauer, S. J. (2011). Taking the reins: The effects of new leader status and leadership style on team performance. *Journal of Applied Psychology*, *96*(3), 574–587.
Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038.
Shively, M. (2001). Male self-control and sexual aggression. *Deviant Behaviour*, *22*(4), 295–321.
Skinner, C. J. (1986). Design effects of two-stage sampling. *Journal of the Royal Statistical Society: Series B*, *48*(1), 89–99.
Smith, B. (2018). Generalizability in qualitative research: Misunderstandings, opportunities and recommendations for the sport and exercise sciences. *Qualitative Research in Sport, Exercise and Health*, *10*(1), 137–149.
St John, C., & Heald-Moore, T. (1995). Fear of black strangers. *Social Science Research*, *24*(3), 262–280.
Sticca, F., & Perren, S. (2013). Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of Youth and Adolescence*, *42*(5), 739–750.
Taylor, B. J. (2005). Factorial surveys: Using vignettes to study professional judgement. *British Journal of Social Work*, *36*(7), 1187–1207.
Taylor, B. J., & Zeller, R. A. (2007). Getting robust and valid data on decision policies: The factorial survey. *The Irish Journal of Psychology*, *28*(1-2), 27–41.
Ventsislavova, P., Crundall, D., Baguley, T., Castro, C., Gugliotta, A., Garcia-Fernandez, P., Zhang, W., Ba, Y., & Li, Q. (2019). A comparison of hazard perception and hazard prediction tests across China, Spain and the UK. *Accident Analysis & Prevention*, *122*, 268–286.
Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, *38*(3), 505–520.
Wason, K. D., Polonsky, M. J., & Hyman, M. R. (2002). Designing vignette studies in marketing. *Australasian Marketing Journal*, *10*(3), 41–58.
Webster, E. S., Paton, L. W., Crampton, P. E. S., & Tiffin, P. A. (2020). Situational judgement test validity for selection: A systematic review and meta-analysis. *Medical Education*, *54*(10), 888–902.
Whetzel, D., Sullivan, T., & McCloy, R. (2020). Situational judgement tests: An overview of development practices and psychometric characteristics. *Personnel Assessment and Decisions*, *6*(1), 1–16. https://doi.org/10.25035/pad.2020.01.001

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Baguley, T., Dunham, G., & Steer, O. (2022). Statistical modelling of vignette data in psychology. *British Journal of Psychology*, *113*, 1143–1163. https://doi.org/10.1111/bjop.12577