FORUM

The Use of Questionnaire Data in Presence Studies: Do Not Seriously Likert

Abstract

The problems of valid design of questionnaires and analysis of ordinal response data from questionnaires have had a long history in the psychological and social sciences. Gardner and Martin (2007, this issue) illustrate some of these problems with reference to an earlier paper (Garau, Slater, Pertaub, & Razzaque, 2005) that studied copresence with virtual characters within an immersive virtual environment. Here we review the critique of Gardner and Martin supporting their main arguments. However, we show that their critique could not take into account the historical circumstances of the experiment described in the paper, and moreover that a reanalysis using more appropriate statistical methods does not result in conclusions that are different from those reported in the original paper. We go on to argue that in general such questionnaire data is treated far too seriously, and that a different paradigm is needed for presence research—one where multivariate physiological and behavioral data is used alongside subjective and questionnaire data, with the latter not having any specially privileged role.

1 Introduction

The vast majority of experimental studies of presence in virtual environments have used questionnaires to assess presence and related constructs. Questionnaires inevitably result in ordinal response data, and these are typically analyzed using some variant of standard Analysis of Variance (assuming a normal error distribution), in order to test for differences in means across factor levels. One such study was reported in Garau, Slater, Pertaub, and Razzaque (2005), and this has been used as an exemplar by Gardner and Martin (2007) in order to illustrate how *not* to pose questions in questionnaires, and how *not* to do the statistical analysis. As was pointed out by Gardner and Martin, the paper by Garau et al. is far from unique with respect to the analysis methods chosen—these being all too common in the psychological and social sciences. In this paper we address the comments made by Gardner and Martin with respect to the points they raise concerning appropriate wording of questionnaires and the analysis of ordinal response data, together with comments about the interpretation of statistical results in the general context of scientific method. We root our discussion primarily in the context of studies of presence and the utility of sole reliance on questionnaires in this context. In the remainder of this paper we refer to the paper of Gardner and Martin as GM, and the paper by Garau et al. as GSPR.

2 Question Structure and Likert Lumpiness

The comments of GM regarding the wording of questions and the possibility of introducing bias in answers as a result are unassailable. In particular they note that when a question is phrased so that the answer might nor-

Presence, *Vol. 16, No. 4, August 2007, 447–456*

Mel Slater*
Institució Catalana de Recerca i Estudis Avançats (ICREA)
Universitat Politècnica de Catalunya
Centre de Realitat Virtual
Edificio U, C. Pau Gargallo, 16
08028 Barcelona, Spain
and
Department of Computer Science
University College London
Gower Street
London, WC1E 6BT UK
Maia Garau
Tiny Pictures Inc.
San Francisco, CA 94103

*Correspondence to melslater@lsi.upc.edu

mally be binary, but where the response available is a 7-point scale, then the results might tend to be biased towards the extremes. For example, if you ask a number of people: “Do you prefer ice cream or chocolate?” and the possible answers are: 1 (ice cream) . . . 7 (chocolate)—the implication might be that the frequency response is likely to be a U-shaped distribution since the question is binary.

In the GSPR experiment there were 41 subjects, who had an experience within a Cave-like immersive virtual environment in a between-group experiment with 4 factors. The participants walked through a virtual library that contained 5 virtual characters (‘avatars’). In condition 1 the avatars were static. In condition 2 they were animated—such as fidgeting, turning pages of a book and so on—but paid no attention at all to the participant. In condition 3 there was a spatially based response to the participant—that is, an avatar would turn and look at the participant when within a certain distance, and the facial expression would change. Finally condition 4 was the same as 3 but where the avatar initiated some verbal interaction. After their experience the participants completed a questionnaire, and we include here some of the questions of the type that ought to result in such lumpy data, set out in the way that they were in the actual questionnaire:

A. During the course of the experience, which was strongest on the whole, your sense of being in the room, or of being in the real world of the laboratory?

I had a stronger sense of . . .								
Being in the lab	1	2	3	4	5	6	7	Being in the room

B. During the time of the experience, did you often think to yourself that you were just standing in a laboratory or did the room overwhelm you?

During the experience I was thinking that I was really in laboratory . . .								
Most of the time	1	2	3	4	5	6	7	Rarely

C. During the course of the experience, did you have a sense that you were in the room with other people or did you have a sense of being alone?

With other people	1	2	3	4	5	6	7	Alone

GM also argue that questions of the following type might result in responses at the upper extreme:

D. Please rate *your sense of being in the* room, on the following scale from 1 to 7, where 7 represents your *normal experience of being in a place.*

I had a sense of being there in the room . . .								
Not at all	1	2	3	4	5	6	7	Very much so

E. During the course of the experience, how much were you aware of the experimenters?

Not at all	1	2	3	4	5	6	7	Very much

F. How aware were you of the characters in the room?

Not at all	1	2	3	4	5	6	7	Very much

G. To what extent did you have a sense of being in the same space as the characters?

Not at all	1	2	3	4	5	6	7	Very much

In the case of questions D–G (there are other examples in the questionnaire—too many to reproduce here) GM argue that the problem is not in the wording of the question but in the use of the term “Very Much” to describe the upper extreme of the scale. The issue here is that there may be respondents for whom “Very Much” could be insufficient, and therefore there could be a concentration of answers in that category since it

Table 1. *Frequency Distributions of Responses to Questions A–G, n = 41*

Score	A	B	C	D	E	F	G
1	2	2	2	0	15	1	1
2	5	5	2	1	13	2	2
3	2	5	10	6	5	5	4
4	5	4	4	7	4	2	1
5	9	9	2	16	3	5	10
6	14	9	9	6	0	11	13
7	4	7	12	5	1	15	10

would include all the "very much," "really very much," "completely," and so-on types of responses; very much means different things to different people.

Table 1 shows the frequency distributions for these questions. Recall that A–C should have U-shaped distributions, since they are binary and hence should have answers biased towards the two extremes, whereas D–G should be biased towards a concentration of answers in the uppermost category. The data do not show distributions that might be expected from the GM arguments, especially A–C do not resemble U-shaped distributions, nor is there any evidence in D–G that the Very Much category is treated in a particularly special way.

Now it could be argued that we cannot know the influence of the form of question nor category terminology: had the question been phrased in a different way, or had the category descriptions been different, then the answers may have been different. This is true of course, and emphasizes the real point of GM that questions must be phrased very carefully in order to avoid ambiguities, or biases. However, there are two points that the argument does not take into account. The first is the fact that the questions are embedded in a whole questionnaire that provides context for the specific questions identified as problematic. In other words respondents are to some extent trained by the totality of the questionnaire regarding the appropriate way to interpret these types of questions. Consider, for example, the very first question:

H. Please rate the extent to which you were aware of background sounds in the laboratory in which this experience was actually taking place. Rate this on the following scale from 1 to 7 (where for example 1 means that you were not at all aware of the background sounds):

During the experience I was aware of background sounds from the laboratory . . .

Not at all 1 2 3 4 5 6 7 Very much so

It seems reasonable to suppose that participants who were unaware of background sounds would answer 1, those who were always aware of background sounds would answer 7, and those with experiences between these extremes would answer somewhere between these two (as it happens 78% answered with scores 1 or 2).

The second question was:

I. How dizzy, sick or nauseous did you feel resulting from the experience, if at all? Please answer on the following 1 to 7 scale.

I felt sick or dizzy or nauseous during or as a result of the experience . . .

Not at all 1 2 3 4 5 6 7 Very much so

The same argument can be applied—this is a question that asks people to report directly their experience in an area that needs no special understanding or training. The questions relate to something that people can report on directly (did they hear sounds? did they become sick?), and in everyday conversation one could ask these types of questions (leaving out the numerical scale) and a person might respond: "I wasn't sick at all" or "I felt a bit sick, but not much" or "I was really very sick" and so on. Of course, one could ask such questions in an open-ended way and get responses such as these—but the temptation would be overwhelming to then rank all the responses in an ordering from lowest level of sickness to highest—in order to obtain ordinal data. When people are making these judgments of course they are basing their choice of words (or score) on their own personal history of feeling sick. It is exactly these comparisons that we need. (As a matter of interest 60% answered scores 1 or 2 on question I.)

Of course there is no pinpoint accuracy here—whether the answer is 5 or 6 on the one hand or 2 or 3 on the other really does not matter—the analysis is interested always in trends, and the numerical scores are simply a way of making life easy, a way to avoid dealing with masses of textual data.

We suggest that mixing the questions of interest (e.g., about presence) between other more straightforward questions that relate to people's normal everyday experience (e.g., how much they were aware of noises, how sick they got, how much they use a computer in their everyday work, and others) acts as a kind of training for how to approach the questions about more abstract aspects of experiences (such as presence).

The second point is that it is very hard to believe that participants read these questions with an analytical frame of mind, paying attention to the exact form of wording used. It is true that it is vital to avoid ambiguity, but in our experience when questions are ambiguous the participants tell us! We always pilot experiments extensively, usually over several months, with roughly as many pilot subjects as for the real study, and obviously with a cumulative knowledge built up over many years. The questionnaire is a critical aspect of this piloting process—ambiguities and errors are ironed out through discussion with the pilot experiment participants.

Overall then regarding the issue of wording in questionnaires, again we emphasize that we do not disagree with the views of GM. Obviously it is always better to have watertight questions rather than those of arguably lesser quality, and the suggestions they make for improvement should be adopted. However, we argue that this should not be taken too seriously—empirically our data suggests that nothing in particular went wrong. Moreover, questions should not be looked at in isolation but rather the entire questionnaire structure needs to be taken into account holistically. Finally, it is highly unlikely that questionnaire respondents study each question concentrating on the exact phraseology used with an analytic frame of mind—they tend to answer quickly assessing the intent of the question.

The paragraph above makes a number of statements that could themselves be empirically tested by specific studies of questionnaire design. Would we advise that such research be carried out? In the context of presence research we believe that this would be a waste of scarce resources—and explain why later.

3 Ordinal Data in Regression

The strategy that we have adopted in our studies of presence over many years, up until recently, was to carry out between-group experiments. In such experiments critical factors of interest were varied across the groups. In the case of the GSPR paper, the major factor of interest was the impact on the participants of the degree of interactivity portrayed by virtual characters that the experimental participants met in the setting of a virtual library. There were a number of response variables of interest, for example, one that we labeled as copresence. The method of assessing copresence was through questionnaire responses. That paper was our first that marked a new direction where we attempted to examine questionnaire responses in conjunction with physiological and behavioral ones, but here we only focus on the methodology regarding the questionnaires.

Now how do we arrive at a set of questions that describe the meaning that we wish to attribute to copresence? One source of questions is through listening to what participants themselves tell us during debriefing interviews from past experiments and during debriefing interviews of the current experiment. This is somewhat of a bootstrapping exercise—of course we start with some set of questions that we invent ourselves, but these become refined (or dropped altogether) over time. The particular five questions that we used for copresence were given in the original GSPR paper and were distributed throughout the 45-item questionnaire. (Some of the questions have score 1 meaning low copresence and score 7 meaning high copresence and others are reversed. In all that follows it is assumed that scores have been adjusted before use in analysis so that always a higher number means higher copresence.)

In such an experiment it is impossible within available resources to control for every possible confounding factor (even if we knew what they were). For example, differences in response across the conditions might be due to differences in age, gender, how much the participants had experienced virtual reality before, how much they know about computer graphics or virtual reality, their personality type, whether they are taking certain medicines, whether they have recently consumed alcohol, and so on. Instead of trying to match many characteristics between groups, we typically give a prequestionnaire that gathers information on many possible confounding factors. For example, in this particular study we reasoned that the participant responses to the virtual people might be influenced by how socially anxious they might be in everyday life—so we also administered a particular questionnaire prior to the experiment that is supposed to assess this (Watson & Friend, 1969). Finally given the response (dependent) variables of interest (e.g., copresence), the main factors (independent variables), and the explanatory variables (possible confounds), we carry out an Analysis of Covariance. On the left-hand side of the model is the dependent variable, and on the right-hand side appropriate linear combinations of the independent and explanatory variables. Note that all techniques that are typically used in this context are special cases of the General Linear Model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is an $n \times 1$ vector of responses (in this case $n = 41$), $\mathbf{X}$ is $n \times k$ matrix that is made up through the combination of independent and explanatory variables, with typically the first column consisting only of 1s to allow for a general intercept term (or grand mean). The usual tests of statistical inference assume that $\boldsymbol{\epsilon}$ is a vector of n independently and normally distributed random variables, with constant variance. The $k \times 1$ parameter vector $\boldsymbol{\beta}$ is estimated through least squares, and hypothesis tests on the parameters and tests of goodness of fit of the model as a whole rest on the normality assumption. In particular, a null hypothesis is that $\beta_j = 0$ for $j = 2, 3, \ldots, k$, which is equivalent to the statement that none of the independent or explanatory variables have an influence on $\mathbf{y}$.

Now of course there is a problem. Much of the data is ordinal as obtained from the questionnaire responses. We distinguish between two cases—ordinal variables on the right-hand side (explanatory variables) and on the left-hand side (dependent variable). There isn't really too much of a problem with ordinal variables on the right-hand side of the equation—as GM note, most researchers would turn a blind eye to summing ordinal variables across individuals: recall that we are interested in trends not exactitude in this context. There is, however, a problem as pointed out by GM in using ordinal response variables on the left-hand side. These are highly unlikely to satisfy the normality assumptions discussed above. For any single ordinal response variable one solution is to step up to a higher order model, the generalized linear model (McCullagh & Nelder, 1989; of which the general linear model is a special case). This does not require normality, but that the distribution of the response is a member of the exponential family (this includes many well-known distributions such as binomial, Poisson, gamma, and of course normal, as special cases). It also does not assume that the expected value of the response variable is equal to the linear combination of X variables, but that this relationship may be mediated through a nonidentity "link function." Using the generalized linear model we have two alternatives when the response variable is ordinal. The first is to collapse the ordinal response variable into a binary one (e.g., on a 7-point Likert scale, all scores of 4 and above might be counted as 1 and all scores below 4 as 0). Then logistic regression may be used, which assumes that the response variable has a binomial distribution. The second alternative would be to use ordinal logistic regression, which does not throw away any of the ordinal data as is inevitably the case with the reduction to binary form (McKelvey & Zavoina, 1994; O'Connell, 2005).

The real problem then is not dealing with an individual ordinal response variable, but in combining several ordinal variables together to produce one overall variable (copresence in this case). GM point out the many problems in using averages of ordinal variables as a single response variable, as was done with the copresence construct. Again their arguments are unassailable. Of course we cannot (and did not) assume that taking the average across 5 obviously correlated ordinal responses would bring the Central Limit Theorem into play and result in normally distributed response. In fact it is not the response itself that needs to be normally distributed with zero mean and constant variance but the response under the null hypothesis (of no relationship between **y** and the columns of **X**). Specifically it is the residual errors of the model that need to satisfy the normality assumption. If we look at the empirically obtained residual errors of the model used in the GSPR paper, we find that the normality assumption for their distribution is not rejected. For example, a Kolmogorov-Smirnov test does not reject the null hypothesis of normality ($p = .965$) and similarly the Jarque-Bera test of normality does not reject this hypothesis ($p = .64$).

At first sight there is a puzzle. GM are right to say that averaging across such ordinal data may lead to meaningless results, and that the normality assumption would not likely be met. But in fact in our experiment the results were quite meaningful, and the statistical assumptions do not seem to be violated. Perhaps the answer lies in experience. This particular experiment did not come out of the blue. The constructs used and the questions asked were not suddenly invented from out of nothing—but reflected experience from many years of prior research. This is not to say that such techniques of analysis are recommended, only that an abstract critique may not appropriately predict likely outcomes without taking into account historical context—both with respect to this particular experiment, and all those that went before it and led up to it.

4 Reanalysis

4.1 Logistic Regression

However, we did, in that particular paper, use averages across ordinal responses. In almost every other experiment we have avoided this for exactly the reasons stated by GM. The technique we have used before is different and more conservative (e.g., Slater, Steed, McCarthy, & Maringelli, 1998) and although not perfect nevertheless does not make the authors so statistically uncomfortable. Suppose we have N ordinal response variables ($N = 5$ for copresence above) $y_1, y_2, \ldots, y_N$ each on a 1–7 Likert scale. Then we form a new variable Υ which is the number of y_j such that $y_j \geq K$, $j = 1, 2, \ldots, N$. The new variable Υ represents the number of high responses among the N questions. The meaning of "high" depends on the value of K, for example $K = 5$ or $K = 6$ are typical choices. Here we can take one of the arguments of GM to an extreme and suppose that the experimental participants assign scores to the questions at random. We can allow different patterns of randomness for each participant—in particular we assume that participant i assigns scores randomly such that the probability of a high score (in the sense defined above) is p_i, and that the assignments of the scores over questions are independent for any one subject, and of course independent across subjects.

Now let Υ_i be this overall response variable for the ith subject. Under the assumptions above the distribution of each Υ_i is independently binomial with parameters N and p_i. In analysis we can therefore use logistic regression, one of the variants of the generalized linear model. This model is of the following form:

$$E\left(\frac{\Upsilon_i}{N}\right) = p_i = \frac{1}{1 + \exp(-\eta_i)} \qquad (1)$$

where

$$\eta_i = \sum_{j=1}^{k} \beta_j x_{ij}$$

The particular link here between the expected value of the response variable and the linear component η is the logistic function, and readers are referred to statistical texts to obtain in-depth explanations of this (McCullagh & Zavoina, 1989). As before the hypothesis of interest is that $\beta_2 = \beta_3 = \cdots = \beta_k = 0$, which if rejected suggests that at least one of the X variables covaries with Υ. In practice there are methods for examining the influence of variables individually, the effect of adding or deleting variables into the model, and an assessment of the overall fit of the model. If the null hypothesis is not rejected, it suggests that each participant assigned

Table 2. *Principal Components of the Five Copresence Question Responses Ordered from Highest to Lowest Contribution to the Total Variance**

Variable	PC1	PC2	PC3	PC4	PC5
y_1	0.3753	0.8139	0.4027	0.0461	0.1800
y_2	0.4862	–0.2618	0.4061	–0.3082	–0.6597
y_3	0.2617	–0.3166	0.2634	0.8697	0.0743
y_4	0.5176	–0.3696	–0.0034	–0.3481	0.6887
y_5	0.5352	0.1793	–0.7769	0.1592	–0.2293
Variance	7.5793	3.4920	2.0672	1.9320	1.2234
%	46.5	21.4	12.7	11.9	7.5

* The 5 × 5 matrix of PC coefficients is the matrix of eigenvectors of the covariance matrix of the original set of covariance scores.

scores at random, the only difference between them being that they have different propensities to assigning high scores— but that their assignments are not related to the X variables. When we carried out this analysis on the original data, we obtain the same results on all the regression analyses as given in the GSPR paper. In particular, it was found that condition 3 (responsive avatars) resulted in the highest copresence, but that there was an interaction effect with the degree to which people use computers in everyday life, such that the greater their reported computer usage the lower the reported copresence—the same result is reproduced with the logistic regression analysis as with the ordinary regression analysis.

This approach avoids treating ordinal response data as if it were interval, and by construction of the new response variable, the distribution under the null hypothesis is binomial. There remains an uncomfortable issue regarding the assumption of independence between the questions. In general each circumstance has to be argued on its merits—whether the different dimensions represented by the different questions can be argued to produce independent responses, and this may also be helped by scattering the relevant questions throughout the questionnaire. The problem is that if they are not independent then the N referred to above is no longer valid, so that the significance levels obtained could be inflated. In fact when we examine the correlations between the responses to the five copresence questions, some of them are highly correlated. We address this in the next section.

4.2 Principal Components in Logistic Regression

We carry out a Principal Components Analysis (PCA) on the raw copresence data. This results in five new variables (the principal components—PCs) that are linear combinations of the original variables, but which are orthogonal (i.e., uncorrelated). The fact that they are uncorrelated does not imply that they are independent unless the PCs are also from a multivariate normal distribution— but it is the best we can do.

Table 2 shows the principal components of the copresence questions and their variances. For example

$$PC_1 = 0.3753y_1 + 0.4862y_2 + 0.2617y_3 + 0.5176y_4 + 0.5352y_5$$

and var(PC_1) = 7.5793, which is 46.5% of the total variance. The total variance of all the PCs is the same as the total variance in the original set of question responses, by construction of the principal components. In other words we have five new variables but where the new variables are uncorrelated with one another, and where PC_1 has the highest variance, PC_2 the next highest, and so on.

Let **Y** be the 41 × 5 matrix of the original copresence scores, **P** the 41 × 5 matrix of corresponding principal component scores, and **V** the 5 × 5 matrix of principal component coefficients (in fact the eigenvectors of the covariance matrix of **Y**), which is the matrix shown in Table 2. Then by construction of the principal components, it is easy to show that $\mathbf{Y} = \mathbf{PV}$ and hence $\mathbf{P} = \mathbf{YV}^T$ since **V** is orthogonal. Hence given any set of values of $y_1, y_2, \ldots, y_5$ we can find the corresponding values of $PC_1, PC_2, \ldots, PC_5$. Therefore it is easy to find the range of PC scores corresponding to the situation where *all* $y_j > 3$, $j = 1, 2, \ldots, 5$ for example. This gives us the range of PC scores corresponding to the situation where *all* copresence questions resulted in scores greater than 3. For example all $y_j > 3$ corresponds to $PC_1, PC_2, \ldots, PC_5 >$ 7.6156, 0.1585, 1.0220, 1.4652, 0.1891, respectively. Now we adopt the strategy of counting the number (Υ_{PC}) of PC scores out of $N = 5$ that are greater than these limits and use this as the binomial response variable, but in this case with five uncorrelated variables. Using this alternative construct for copresence the logistic regression can be carried out now using Υ_{PC} as the response with $N = 5$. The results are almost the same as in the previous analysis. This time conditions 1 and 2 are not significantly different from one another, and conditions 3 and 4 are not significantly different from each other but are significantly higher than 1 or 2 (all at the 5% level). Moreover, as before the variable computer (the extent to which participants report using a computer in their everyday life) is negatively associated with copresence, but this time there is no interaction effect—it is the same for all four conditions. We would conclude that the evidence suggests that there is an impact of the avatar behavior on this copresence construct, and that the conditions where the avatars pay attention to the participants results in higher copresence than the conditions where they do not. Moreover the animated avatars that pay no attention to the participants appear to be no better than static avatars. The results are summarized in Table 3.

Table 3. *Estimates and Standard Errors PCA-Copresence Score Logistic Regression*

Term	Estimate	Standard error
General mean (μ)	2.295	1.207
α_2-animated	0.626	0.403
α_3-responsive	0.868	0.407
α_4-talking	1.323	0.431
x-computer	–0.390	0.185

The particular form of the linear model from Eq. (1) is $\eta_{ij} = \mu + \alpha_i + \beta \cdot x_{ij}$ where μ is the general mean, α_i is the *i*th condition effect, and x_{ij} is the score on computer usage of the *j*th person in the *i*th condition. Conventionally $\alpha_1 = 0$ in order to obtain a nonsingular covariance matrix. Hence the general mean accounts for condition 1, and the estimates and standard errors are *differences* from the general mean.

The deviance of this model is 46.749 on 36 d.f., which is a fair overall fit, within the lowest 90% of the chi-squared distribution (p = .11). If the condition effect (i.e., the α term) is deleted then the deviance increases by 10.71 on 3 d.f. (p = .0134) and if the computer usage (x term) is deleted then the deviance increases by 4.812 on 1 d.f. (p = .0283). Hence neither term can be deleted from this model without significantly worsening the overall fit.

4.3 Weighted Average

GM also consider the following issue—why in the original paper did we weight the five components of copresence equally—perhaps some components are more important than others? The main answer to this is conceptual: this average of the five questions was, from the point of view of that paper, our operational definition of copresence. We asked the following question—if we define (in an operational sense) copresence to be this particular numerical quantity, what results do we obtain from the experimental data? Above we have operationally defined copresence as a different weighted average—this would be another definition—we are free to define this concept as we like. Other researchers may or may not be interested in how we define it, but it cannot be said a priori that it is incorrect, that there should be different weightings, since those different weightings would provide yet another operational definition of the concept—with no greater epistemological status than the first. Nevertheless since we already have the PCs we can consider this. If we look at the column of coefficients of PC_1 in Table 2, we see that they are in quite a narrow range, and there is no reason to say that any of them is close enough to zero to be eliminated. If we carry out a normal regression with this PC as the response variable (in other words we have a weighted average of five components) we obtain the same qualitatively significant results as in the original analysis (which had equal weights for each component). The weights make no difference—which is not surprising given that the weights do not differ greatly from one another.

4.4 Conclusion

The conclusion to this section that discusses the problem of combining ordinal data from several questions into one overall score, is that in this particular experiment we obtain consistent results from several different approaches. We do not in general support averaging ordinal data from several questions, and GM were right to point out the problems in doing this. We have found that in this particular case it has made little difference. This may be because the questions used were well-founded, based on extensive prior experience. In general, however, we would suggest not to take such Likert scores too seriously in any case—their major role should be in providing support for other more physically and behaviorally based variables in a multivariable context. We take this more general discussion up in the next section.

5 Presence, Questionnaires, and Statistical Significance

We have pointed out before the methodological problems inherent in attempting to provide a measure of presence (or equally copresence) solely by the use of questionnaires (Slater, 2004). The definition of presence that we work with now is different from notions about the sense of being there (although it does include that). Presence in our current work is the extent to which participants respond to virtual sensory data as if it were real, where response ranges from unconscious physiological responses, through behavioral responses, through to feelings, emotions and thoughts (Sanchez-Vives & Slater, 2005). We will not expand on this here but point out that such an approach requires data obtained in several different ways—physiological recordings, many different types of behavioral measures, questionnaires, post-experimental interviews, with other possibilities (depending on the application context). What is important is not any of these in isolation, but the associations between them. An example of this approach can be found in Slater et al. (2006).

In looking at associations between such variables the last thing we care about is the exact statistical significance—since anyway such Likert data is simply an indication of tendency rather than having any exact meaning. GM were right to point out the accident of history that led to 5% significance much worshiped in psychological research, being a gold standard here. But who enforces this? Well of course we ourselves do—the reviewers of papers! Why not instead simply quote the significance level, that is, p(rejection of null hypothesis | null hypothesis)? Better still think of this in a Bayesian sense as proportional to the probability of the null hypothesis when a flat prior distribution is used. While psychological research insists on the 5% rule, in the field of statistics itself there has been a paradigm shift where Bayesian methods have been widely accepted, based on the notion of probability as a subjective degree of belief rather than as observable frequencies of events. So even better still, do a full Bayesian analysis and give the a priori and a posteriori probability distributions and let other scientists judge for themselves—rather than follow an automatic rule, reject the null hypothesis if the significance level is more than 5%.

GM suggest a number of ways in which the analysis of ordinal data could be improved—and we have provided illustrations of some other ideas in the previous sections. However, in the context of presence research we would suggest that this is taking these Likert scores far too seriously. Ultimately suppose that the ideal statistical analysis were carried out, one that resulted in a statistical model that was also successful in predicting scores in future studies. What would have been gained?

Only that the model would be able to successfully predict where on a scale of 1 through 7 participants might place their mark. There are no particularly interesting consequences that follow from this. The application of more and more sophisticated analysis techniques for this type of data (in this context) lends it undue importance. We would argue instead to concentrate resources on collection of data across many variables, use statistical methods to try to obtain some idea of the relationships between them and how they vary with experimental conditions, but treat this only as the starting point for the ultimate aim of building models that *explain,* uncovering the *mechanisms* behind the data, leading to fresh insight into how and why people respond as they do to virtual sensory data. No amount of, for example, factor analysis can substitute for the very hard work of such data collection, analysis, and model construction that this involves.

The GM paper was welcome in once again raising issues of appropriate questionnaire design and statistical methods in the context of presence research. The lesson we draw is not so much to change statistical methods or design even more sophisticated questionnaires, but to change the paradigm. The lesson is, in other words, relegate questionnaires and their analysis to a supporting role only, rather than the lead actor in the presence story, and do not take Likert scores so seriously.

Acknowledgments

Mel Slater's contribution to this article is under the Integrated Project PRESENCCIA, funded by the European Sixth Framework Program, Future and Emerging Technologies (FET) Contract Number 27731.

References

Garau, M., Slater, M., Pertaub, D. P., & Razzaque, S. (2005). The responses of people to virtual humans in an immersive virtual environment. *Presence: Teleoperators and Virtual Environments, 14*(1), 104–116.

Gardner, H. J., & Martin, M. A. (2007). Analyzing ordinal scales in studies of virtual environments: Likert or lump it! *Presence: Teleoperators and Virtual Environments, 16*(4), 439–446.

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models, 2nd ed., pp. xix, 511. New York: Chapman and Hall.

McKelvey, R., & Zavoina, W. (1994). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4,* 103–120.

O'Connell, A. A. (2005). Logistic regression models for ordinal response variables. *Quantitative applications in the social sciences.* Thousand Oaks, CA: Sage Publications.

Sanchez-Vives, M. V., & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience, 6,* 332–339.

Slater, M. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators and Virtual Environments, 13*(4), 484–493.

Slater, M., Steed, A., McCarthy, J., & Maringelli, F. (1998). The influence of body movement on subjective presence in virtual environments. *Human Factors, 40*(3), 469–477.

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., et al. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PLoS ONE* 1 (1), e39. doi: 10.1371/journal.pone.0000039.

Watson, D., & Friend, R. (1969). Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology, 33*(4), 448–457.