Discussion of a Dataset on the Effect of Context on the Speechreading of Spoken Sentences

T.P. Hutchinson and D. Cairns
Macquarie University, Sydney, Australia

Published data from an experiment on the visual perception (speechreading) of spoken sentences, when received as answers in question-answer sequences and when received in isolation, are re-examined. Evidence is found that the extent of the advantage of the question-answer sequence tends to be less for older subjects. As to how to measure the difference in performance in 2 experimental conditions, arguments are given for preferring the difference of logits of the success probabilities over either the difference of probabilities or the ratio of probabilities.

INTRODUCTION

Erber (1992) reported an experiment on the visual perception (speechreading) of spoken sentences, when received as answers in question-answer sequences and when received in isolation. The *subjects* were 24 adults with hearing impairment. The *stimuli* were 40 short sentences, presented on videotape without sound. The *task* was to report the sentence seen. Four words in each sentence (a total of 160 words) were scored as either correct or wrong. The *experimental manipulation* was the context of presentation of the sentences. For each subject, 20 sentences were presented in isolation (condition I). The other 20 sentences were presented in reply to a question (condition Q) – this was given to the subject on a card, the subject read it out, and then watched the video screen for the reply. The experiment was *balanced:* for each sentence, half of the subjects viewed it in isolation

T.P. Hutchinson and D. Cairns, Department of Psychology.
Correspondence concerning this article should be addressed to T.P. Hutchinson, Department of Psychology, Macquarie University, Sydney, N.S.W. 2109, Australia. Phone: (02) 9850 8013. Fax: (02) 9850 8062. Electronic mail may be sent via Internet to phutchin@bunyip.bhs.mq.edu.au.

and half of the subjects viewed it as a reply to a question.

The key finding of Erber's (1992) paper was that, on average, a subject's score was higher for sentences that followed a question than for sentences presented alone (Erber, 1992, Table 1, p. 115; Figure 1, p. 117). This is one feature that can be seen in Figure 1: for the great majority of subjects, p_Q exceeds p_I, where p_Q and p_I are the proportions of words correctly identified in the two conditions (with the sentences being in reply to a question and in isolation, respectively).

Erber (1992) did not attempt to relate the size of the experimental effect – that is, some quantity similar to the difference between the subject's performance in the two experimental conditions – to characteristics of the subject, such as present age and age at onset of impairment. The purpose of the present paper is to take up this issue. But first, it is necessary to specify an appropriate method of measuring the size of the experimental effect. The method to be used is likely to be valid whenever the dependent variable is a probability of success, and thus will be of relevance well beyond the specific context of Erber's experiment.

METHOD

The Defects of the Simplest Approaches

Two possible measures of the effect of the question, as contrasted with the in-isolation condition, are as follows.

- The difference, $p_Q - p_I$. The disadvantage of this is as follows. Suppose it averages 0.20. The implication is that p_Q is never less than 0.20: if it were, p_I would be predicted to be negative, which is impossible. Putting this another way, it seems intuitively reasonable that $p_Q = 0.25$ and $p_I = 0.05$ represents a bigger effect than does $p_Q = 0.50$ and $p_I = 0.30$, even though the difference is 0.20 in both cases.
- The ratio, p_Q/p_I. The disadvantage of this is as follows. Suppose it averages 2.0. The implication is that p_I is never greater than 0.5: if it were, p_Q would exceed 1, which is impossible. Putting this another way, it seems intuitively reasonable that $p_Q = 0.2$ and $p_I = 0.1$ represents a lesser effect than does $p_Q = 0.8$ and $p_I = 0.4$, even though the ratio is 2.0 in both cases.

If the difference were the appropriate measure of the effect of the question, the data in Figure 1 would be expected to fall on a straight line of slope 1 (whose intercept measures the effect of question). On the other hand, if the ratio were the appropriate measure of the effect of the question, the data would be expected to fall on a straight line through the origin (whose slope measures the effect of question). It would be an exaggeration to say that Figure 1 plainly contradicts both these suggestions, but neither does it convincingly support either of them. Thus, although both the difference and the ratio have the advantage of simplicity, they

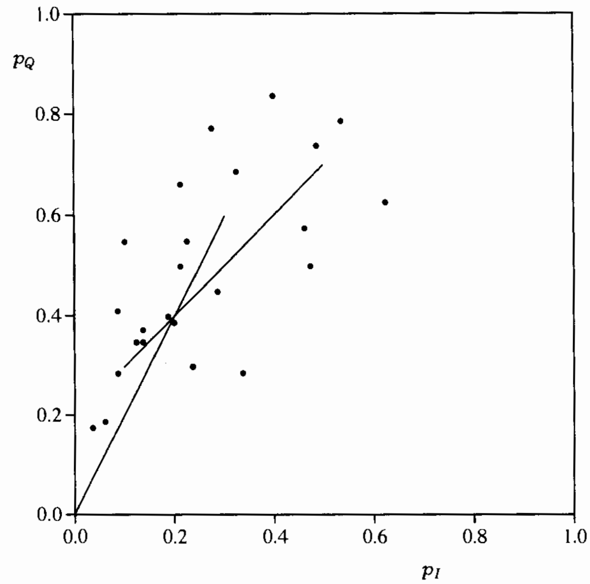

Figure 1. The relation between the probabilities of correctness in two conditions (p_I for the in-isolation condition, p_Q for the question context) for 24 people. A line of slope 1 and a line through the origin are also shown. *Note:* The data in this figure are from "Effects of a question-answer format on visual perception of sentences" by N.P. Erber, 1992, *Journal of the Academy of Rehabilitative Audiology, 25,* p. 115. Copyright 1992 by the Academy of Rehabilitative Audiology.

are unlikely to be satisfactory. Putting this another way, anyone who begins by thinking the data points form more or less a straight line is likely to be convinced otherwise by realizing that the relationship must include the points (0,0) and (1,1).

Logits

There are several methods available to overcome the problems raised, and perhaps the most popular is to use the difference in logits as the measure of effect size. Logit(p) is defined as $\ln[p/(1-p)]$ (where ln means natural logarithm). The difference in logits is $\text{logit}(p_Q) - \text{logit}(p_I)$.

Why the difference in logits? There are two good reasons for a transformation similar to this (but no good reason for exactly this one).

- Statistical models typically involve unknown parameters that combine additively, that is, as in $\alpha_i + \beta_j$. The result may, in principle, be of any size, positive or negative. In contrast, proportions lie in the range 0 to 1. One possible way of getting something within $(-\infty, +\infty)$ from a p

that is within (0, 1) is the transformation $\ln[p/(1-p)]$.

- The logistic distribution is similar in shape to the normal distribution. For the logistic distribution, the formula for the probability p of an observation being greater than t is $p = (1 + e^{t-\mu})^{-1}$, where μ is the mean of the distribution. Rearranging this formula gives $t - \mu = -\ln[p/(1-p)]$. Thus logit(p) can be interpreted as the distance of some threshold from the mean of a distribution. In the context of getting a question correct or wrong, the *threshold* is typically interpreted as the difficulty of the question, and the quantity having a probability distribution is interpreted as a person's ability (conceived of as varying from moment to moment about some mean μ); the question is answered correctly if ability exceeds the difficulty, and this occurs with probability p. For more details about logits, see, for example, section 15.14 of Howell (1997).
- An alternative to the logit is the probit, which is derived from the normal distribution instead of the logistic. The results below are very similar whether logits or probits are used.

RESULTS

The Transformed Graph

If the difference between the logits is indeed the appropriate measure of the effect of the question, a plot of logit(p_Q) versus logit(p_I) would be expected to be a straight line of slope 1. Figure 2 shows that the data support this suggestion. The average difference between the logits turns out to be 1.2. Table 1 gives details of the calculated logits, along with the more important of the independent variables to be used in the regressions below.

To summarize, the starting point was the data in Erber (1992, Table 1, p. 115); the numbers correct were converted to proportions correct, logits were calculated from the proportions, and a straight line of slope 1 has been found. As mentioned earlier, this method may be useful whenever what has been measured is a probability of success.

Regressions

The data are plainly compatible with a slope of 1 when the logits are plotted (see Figure 2). Equally plainly, the points do not exactly fall on a straight line. Is this merely random variation, or is it possible to some extent to explain the variation? The question of whether the difference in performance in the two conditions, logit(p_Q) - logit(p_I), is related to any of the characteristics of the subjects that Erber (1992, Table 1, p. 115) lists will now be investigated. These characteristics were as follows: age, sex, self-reported degree of hearing impairment (mild, moderate, or severe), years of hearing impairment, hearing aids worn

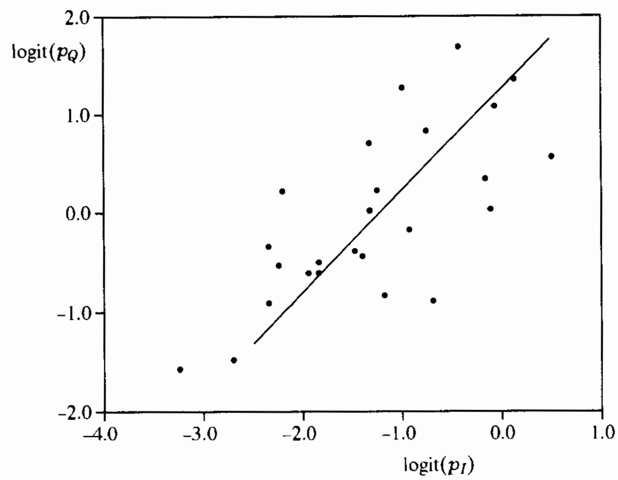

Figure 2. Erber's (1992) data after logit transformation of the probabilities of correctness (p_I for the in-isolation condition, p_Q for the question context) for 24 people. A line of slope 1 is also shown. *Note:* The data in this figure are from "Effects of a question-answer format on visual perception of sentences" by N.P. Erber, 1992, *Journal of the Academy of Rehabilitative Audiology, 25,* p. 115. Copyright 1992 by the Academy of Rehabilitative Audiology.

(none, monaural, or binaural). As well as the independent variables just listed, age at onset of hearing impairment was also included, this being calculated as the difference between age and years of hearing impairment. (The interrelationship of these three variables will be discussed further below.)

It is convenient to report the results using an "elaboration" approach. That is, the equation at first has only one independent variable, and then further independent variables are progressively added, one at a time, into it. See, for example, Bryman and Cramer (1997, chapter 10) for more on this.

1. Is the difference in performance related to any of the independent variables individually? In particular, is age (which is so often an independent variable in studies of hearing) related to difference in performance?
2. Is age still a useful predictor after each of the other independent variables is included also?
3. Is age still a useful predictor after all of the other independent variables are included?

In the regressions to be reported, degree of hearing impairment was treated as a numeric variable, not as categorical, and so was the number of hearing aids worn.

Now to answer the questions raised.

Table 1

Some Characteristics of the Subjects, and Logits of Proportions of Words Correct in the In-Isolation and Question Conditions (p_I and p_Q)

x_1	x_2	x_3	Logit(p_I)	Logit(p_Q)
48	28	20	-0.41	1.64
59	7	52	0.15	1.31
68	12	56	-0.05	1.03
59	48	11	-2.34	-0.91
64	22	42	-0.15	0.30
70	23	47	-0.91	-0.20
79	68	11	-1.31	0.00
66	50	16	-1.39	-0.46
38	36	2	-0.97	1.24
73	35	38	-0.67	-0.91
76	67	9	-1.17	-0.85
52	45	7	-1.84	-0.51
69	44	25	-2.34	-0.35
54	51	3	-1.24	0.20
76	67	9	-1.95	-0.62
62	35	27	0.51	0.51
68	61	7	-2.20	0.20
64	50	14	-1.84	-0.62
79	67	12	-0.10	0.00
76	68	8	-1.47	-0.41
71	60	11	-3.25	-1.55
61	54	7	-2.71	-1.47
71	60	11	-1.31	0.67
60	40	20	-0.73	0.79

Note. x_1 = Age; x_2 = Age at onset; x_3 = Years of impairment.

1. Concerning the correlations of the difference in performance with the various possible explanatory variables, these were mostly small. The largest in absolute magnitude was that with age, which was -0.42. And age was indeed a significant predictor (at the 5% significance level) of difference in performance. The regression coefficient was negative, meaning that increasing age leads to a reduction of the advantage that the context condition has over the in-isolation condition.
2. When other variables were included as well as age in the prediction equation, age remained a useful predictor. (In some cases, it was significant using the 5% significance level, but with other variables only the 10% significance level was attained.) The other variables were not significant.
3. With all the variables included, age was the only significant predictor. Its regression coefficient remained negative.

Comment on Choice of Variables

As mentioned above, age at onset = age - years of hearing impairment. Notice that if one attempts to include all three of these variables in a linear regression, the result is indeterminate – any one of the variables can be replaced by a combination of the other two. This is well known in epidemiology (e.g., Clayton & Hills, 1993, section 31.4). The researcher is able to choose which two of the variables to include on grounds on interpretability. (But another researcher might make a different choice!)

It has already been mentioned that most of the independent variables seem to be unrelated to the difference in performance. The ones that might be important are the trio of age, age at onset, and years of impairment, which are given in Table 1. As to equations involving these, age at onset and present age are a credible and comprehensible pair; the regression equation is

$$y = 3.2 - 0.04x_1 + 0.02x_2 \qquad (1)$$

(where y is difference in performance, x_1 is age, and x_2 is age at onset). Table 2 gives the standard errors of the estimated coefficients. The alternative forms of this equation are

$$y = 3.2 - 0.03x_2 - 0.04x_3 \qquad (2)$$

(where x_3 is years of impairment) and

$$y = 3.2 - 0.03x_1 - 0.02x_3 \qquad (3)$$

Controlling for age at onset, age – or, equivalently, years of impairment – is statistically significant, see Equations 1 and 2. In fairness, another interpretation should be mentioned. Equation 3 is the one that uses the variables as they appear in Erber's (1992) dataset; it so happens that neither of the regression coefficients is statistically significant; so it could be said that no relationship has been found. However, it is the view of the present writers that once either x_1 or x_3 is in the equation, then the appropriate other variable is x_2, because it does not change as time passes.

Table 2

Standard Errors (Shown in Parentheses) of the Coefficients in the Regression Equations

$y = 3.21\ (0.83) - 0.042\ (0.014)x_1 + 0.015\ (0.008)x_2$	(1)
$y = 3.21\ (0.83) - 0.026\ (0.013)x_2 - 0.042\ (0.014)x_3$	(2)
$y = 3.21\ (0.83) - 0.026\ (0.013)x_1 - 0.015\ (0.008)x_3$	(3)
$\text{logit}(p_Q) = 2.25\ (0.98) - 0.021\ (0.017)x_1 - 0.020\ (0.010)x_2$	(4)
$\text{logit}(p_I) = -0.96\ (1.08) + 0.021\ (0.019)x_1 - 0.036\ (0.011)x_2$	(5)

The Contributions to the Difference

Having come thus far in studying the difference in performance, it seems natural to change focus and ask whether the significant relationship with age is due to worsening performance in the context condition, or to improving performance in the in-isolation condition. For this dataset, the answer seems to be both: for performance in the context condition, the regression equation turned out to be

$$\text{logit}(p_Q) = 2.2 - 0.02x_1 - 0.02x_2, \quad (4)$$

and in the in-isolation condition, the regression equation was

$$\text{logit}(p_I) = -1.0 + 0.02x_1 - 0.04x_2, \quad (5)$$

though in neither case was the coefficient of x_1 statistically significant. (Incidentally, this highlights a problem with placing too much importance on "statistical significance": having found x_1 to be nonsignificant in predicting two measures of performance, it would be only natural to jump to the conclusion that x_1 is also nonsignificant in predicting the difference between the two measures, but in fact it is significant.) See Table 2 for the standard errors of the estimated coefficients.

DISCUSSION

Erber's (1992) paper was focussed on the possible difference between levels of performance in two conditions. He established that one existed. The present paper has proposed how it may best be measured. And it has been found that age is a significant explanatory variable: it seems that, when age at onset of impairment is controlled for, increasing age (or years of impairment) leads to a reduction of the advantage that the context condition has over the in-isolation condition.

The analyses in both Erber (1992) and the present paper are limited by the data being in aggregated form. That is, for each subject, only the numbers of words correct in the I condition and in the Q condition are given. As an alternative, suppose that for each *combination* of subject and word, the word condition (I or Q) and the response (correct or wrong) were given. Specialists in educational measurement have developed models appropriate for disaggregated data. Let b_i = the difficulty of item i, θ_j = the ability of subject j, and p_{ij} = the probability of subject j responding correctly to item i. One possible model is logit $(p_{ij}) = \theta_j - b_i$. This could be fitted to the data from the two conditions, and conclusions drawn from comparing the estimated values of the parameters in the two conditions. See, for example, Hambleton and Swaminathan (1985) for more on item response theory, as this subject is called.

Conclusions can only be tentative: Erber's (1992) subjects were only 24 in number and seem to have been a sample of convenience, many of the results obtained have been just significant or just non-significant, and the present authors are not sufficiently knowledgeable about hearing to impose an expert view on the data – rather, this paper has followed fairly passively where the data have led us. Nevertheless, it is hoped that the above comments on Erber's data have been methodologically helpful and substantively suggestive.

REFERENCES

Bryman, A., & Cramer, D. (1997). *Quantitative data analysis with SPSS for WINDOWS. A guide for social scientists.* London: Routledge.

Clayton, D., & Hills, M. (1993). *Statistical models in epidemiology.* Oxford, England: Oxford University Press.

Erber, N.P (1992). Effects of a question-answer format on visual perception of sentences. *Journal of the Academy of Rehabilitative Audiology, 25,* 113-122.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Howell, D.C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury.