Individual differences in susceptibility to the "irrelevant speech effect"

Wolfgang Ellermeier[a] and Karin Zimmer
Institut für Psychologie der Universität Regensburg, 93040 Regensburg, Germany

(Received 20 November 1996; revised 4 April 1997; accepted 11 June 1997)

Individual differences in objective effects of noise on performance were analyzed with respect to their distribution, temporal stability, and the precision of measurement to be attained. Seventy-two subjects had to memorize sequences of visually presented digits while being exposed to one of three auditory background conditions which were randomly mixed on a trial-by-trial basis: (1) foreign speech; (2) pink noise; and (3) silence. Individual ''irrelevant speech effects,'' operationalized by the difference in recall errors under speech and in silence, were normally distributed over a wide range extending from slight facilitation to severe disruption. When 25 subjects repeated the experiment after four weeks, the individual differences were replicated with a reliability of $r_{tt} = 0.45$. Internal consistency, a measure of the precision with which individual effects can be measured in a single session, was moderate ($\alpha = 0.55$). However, both retest, and consistency coefficients are severely attenuated by the use of (sound-minus-silence) difference scores, the reliability of which is bound to be considerably lower than that of the original error scores whenever these are correlated. Given that the original error rates in a specific auditory condition can be determined with reliabilities approaching 0.85, it may be concluded that individual performance decrements due to noise can be reliably measured in the ''irrelevant speech'' paradigm. Self-report measures of noise susceptibility collected to explore potential sources of the large inter-individual variation exhibited only weak relationships with the objectively measured noise effects: Subjects were quite inaccurate in assessing their individual impairment in the three auditory conditions, and a questionnaire-based measure of general noise sensitivity only accounted for a small portion of the variance in objectively measured performance decrements, although in both cases the predictive relationship was much stronger in female than in male subjects. [S0001-4966(97)00110-0]

PACS numbers: 43.50.Qp, 43.72.Dv [GAD]

INTRODUCTION

In environmental noise research, the need to study individual differences has always been more apparent than in other areas of psychoacoustics. One obvious reason is that much of the survey research concerned with noise evaluation uses correlational statistics. Individual differences on a subjective dimension (e.g., annoyance) are correlated with other subjective (e.g., attitudinal) or objective measures (such as exposure levels) characterizing individuals or groups of respondents. Appreciable individual variation and its reliable measurement are crucial to this research approach.

Therefore, instruments for measuring individual differences in annoyance with various noise sources (Job, 1988; Taylor, 1984, for reviews), in response criteria for reporting distress (Green and Fidell, 1991), or in general noise sensitivity (Weinstein, 1978) have been developed. Nevertheless, a recent review (Staples, 1996) blamed a lack of understanding of individual differences in reaction to noise for costly policy failures in the implementation of noise abatement, or traffic rerouting programs.

In contrast to the questionnaire-based noise evaluation studies thus characterized, research into the objective effects of noise on performance—focusing on the demonstration of overall effects of experimental manipulations—has shown little concern with individual differences (Jones and Davies, 1984, for an earlier review). There are occasional reports of personality variables such as anxiety or extraversion interacting with noise effects (summarized in Smith and Jones, 1992); analyses of the stability of individual differences, however, turn out to be rather disconcerting. Smith *et al.* (1981), for example, in an experiment, which required subjects to memorize lists of words both in the quiet and under continuous white noise, found quiet-noise differences in recall scores, and in indices of higher-order cognitive processing (''clustering'') to produce correlations as low as $r = 0.05$ between two sessions one week apart. That is, subjects appearing particularly susceptible to noise in the first session were not the same ones as those showing the largest performance decrements in the second session, and the magnitude of the correlation indicated almost nonexistent individual stability of these noise effects.

This inconsistency of performance across sessions may stem in part from the highly variable effects of continuous or intermittent white noise. Generally, these earlier studies show that white noise presented at high sound-pressure levels may either improve, depress, or result in no change in performance. Moreover, those factors that predict such outcomes cannot be articulated with any degree of certainty. Such inconsistency in mean effects suggests (but is by no

[a]Electronic mail: wolfgang.ellermeier@psychologie.uni-regensburg.de

means definitive in suggesting) that reliability measures may be poor in such settings.

While the studies thus characterized all employed broadband noise of relatively high level ($\geqslant$80 dB), more recent research, pioneered by Colle and Welsh (1976) as well as by Salamé and Baddeley (1982), found highly replicable overall noise effects (no single instance of improvement has been encountered, for example) when using temporally structured sounds of moderate intensity, and a particular task requiring recall ''in the correct order.'' The phenomenon referred to has been termed the ''irrelevant speech effect'' (ISE), since the presentation of auditory material (typically speech) which the subject is told to ignore, and which is of no significance to the task performed, has sizable effects on the serial recall of visually presented items such as letters or digits (for reviews see Jones and Morris, 1992; Jones, 1993; Jones *et al.*, 1996). In recent years, the ISE paradigm has become prototypical for studying moderate-level noise effects in a situation representative of modern office environments. The quickly growing number of studies on the effect have either addressed the cognitive mechanisms involved (e.g., Salamé and Baddeley, 1982; Buchner *et al.*, 1996), or the properties of the auditory distractors producing maximal interference (e.g., Jones and Macken, 1995a; Jones *et al.*, submitted; Ellermeier and Hellbrück, in press); individual differences, however, have not been analyzed to our knowledge, and even the presentation of standard errors seems to be the exception rather than the rule (see however, LeCompte, 1994; Jones and Macken, 1995b).

In our opinion, the need for laboratory studies of individual differences in the susceptibility to noise expressed in a recent review (Staples, 1996) is best addressed by looking at the paradigm characterized above, for which there is ample and consistent evidence of performance disruption. Consequently, the present study was designed to collect data on a fairly large number of subjects ($N=72$) under standard ''irrelevant speech'' conditions. More specifically, a foreign language (Japanese) was used to elicit the effect unconfounded by semantic content, and a ''placebo control'' (pink noise) was presented in addition to the quiet baseline, in order to control for unspecific or expectation-based effects due to the mere presence of an acoustical distractor.

The study was conducted with two goals in mind: The primary goal was to assert the presence of individual differences in noise susceptibility in a controlled laboratory experiment, and to show that these can be reliably measured, applying established psychometric methods derived from classical test theory (Nunnally, 1978; Kline, 1993; Lienert and Raatz, 1994). According to this approach, errors in the ''irrelevant speech'' paradigm are treated much like errors in an intelligence test, the precision and replicability of which is to be determined. A secondary goal was to link behavioral effects to differences in self-reported noise sensitivity in order to explore (a) if subjects are able to assess their susceptibility to a given noise, and (b) if the personal attribute of ''noise sensitivity'' (Weinstein, 1978) may account for some portion of the variance in error rates observed in the laboratory.

I. GENERAL METHOD

A. Subjects

Seventy-two students at the University of Regensburg (median age 24, range 19–44; 31 male, 41 female) participated as subjects. A random subset of this sample consisting of 25 subjects was asked and agreed to participate in a retest session four weeks later. Hearing problems, knowledge of Japanese, or prior experience in ''irrelevant speech'' experiments were exclusion criteria for the experiment. All subjects were naive both with respect to the literature on noise effects, and to the specific hypotheses being investigated.

B. Apparatus and stimuli

1. Visual stimuli

The visual material to be memorized consisted of random permutations of the digits 1 through 9, presented sequentially in the center of a colour monitor. The digits were about 2 cm in height and appeared for 800 ms each, with 200-ms pauses between digits.

2. Auditory stimuli

The ''irrelevant'' auditory materials were recorded and played with 8-bit resolution and an 22-kHz sampling rate using a ''Soundblaster-compatible'' PC sound card. Two types of auditory materials were used: (1) Japanese speech, and (2) pink noise. The speech sample consisted of a 15-s segment from a lecture given by a male speaker. The noise sample was recorded from a Bruel & Kjaer (type 1405) noise generator. These single tokens of speech and noise were shaped to have smooth onsets and offsets, and to yield A-weighted, energy-equivalent sound-pressure levels, L_{eq}, of 76 dB, as verified by measurements at the headphones using a Cortex Electronic (model MK 1) artificial head system. Due to its continuous and broadband nature, the pink-noise sample appeared louder: The mean computed loudness levels were 44.7 sone for the noise, and 25 sone for the speech sample. The auditory stimuli were presented diotically over headphones (Beyerdynamic DT 550) in a quiet, but not sound-treated laboratory room [ambient A-weighted sound level approximately 40 dB].

C. Procedure

Each trial was initiated by a 2-s visual warning signal (a square of decreasing size cueing the subject to the point of fixation), after which the stream of 9 digits was displayed at a rate of 1 per s. Following a 5-s retention interval, a 3×3 numerical array consisting of the numbers 1 through 9 prompted subjects to enter the correct serial order by sequentially clicking the computer mouse on the digits displayed.

On sound trials, the acoustical background (speech or pink noise) was present both during the encoding and the rehearsal phase, for a total of 14 s. No sound was presented during the self-paced recall period. Subjects were told to ignore the auditory input while quietly rehearsing the number sequence.

In order to be able to measure individual differences unconfounded with practice effects, the auditory conditions

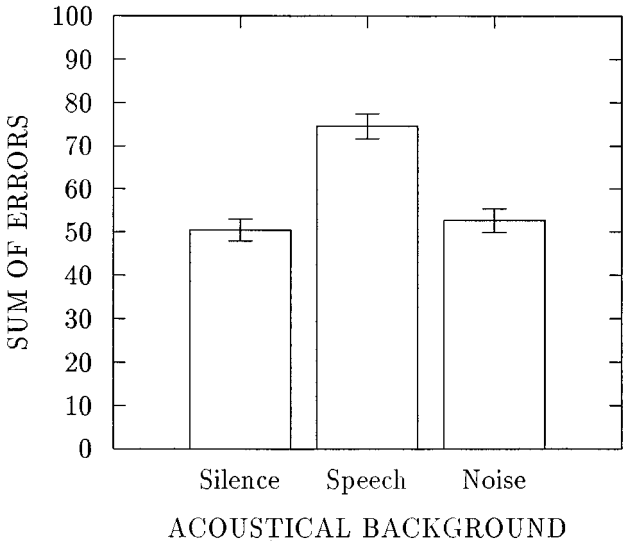

FIG. 1. Effect of ''irrelevant speech'' compared with two control conditions. The sum of serial recall errors in 20 trials averaged across 72 subjects is plotted along with standard errors of the mean.

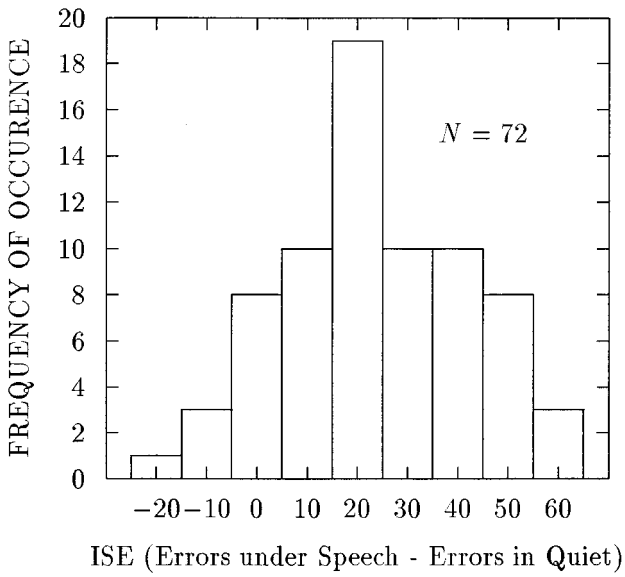

FIG. 2. Distribution of ''irrelevant speech effects'' (ISEs) obtained from 72 subjects. The abscissa shows how far the error rate under speech exceeds the error rate produced in the quiet baseline.

were randomly mixed on a trial-by-trial basis. Subjects were run in blocks of 30 trials in which speech, pink noise, and silent conditions occurred with equal frequency. After three trials of practice, they completed two (or three) of these 30-trial blocks, lasting approximately 15 min each.

Twenty-five subjects repeated the experiment four weeks later in order to determine the stability of the effects over time. Furthermore, this subset of our sample was asked to estimate the degree of interference (or potential facilitation) produced by the ''irrelevant'' sounds by rating them on a bipolar scale ranging from -3 (''will severely hurt my performance'') over zero (no effect) to $+3$ (''will help considerably''). These ratings were obtained after subjects had read the instructions, had heard the two sound samples, but prior to actually performing the serial recall task. The rating procedure was repeated at the end of the first session (with modified wording, where appropriate), in order to assess, whether actual experience with the task changed subjects' evaluation of sound effects.

In addition, all 72 subjects completed two questionnaires measuring individual noise sensitivity with respect to a wide range of noise sources: (1) A German version of Weinstein's (1978) noise sensitivity scale, and (2) a newly constructed noise-sensitivity questionnaire currently being evaluated in our laboratory (Zimmer and Ellermeier, submitted).

II. INDIVIDUAL DIFFERENCES IN SUSCEPTIBILITY TO ''IRRELEVANT SPEECH''

A. Overall effect of irrelevant sound on performance

In all analyses presented in this paper, serial recall performance was evaluated using the method almost exclusively employed in research on ''irrelevant speech'' effects: An error was scored whenever the subject failed to report the correct digit in the correct position. The sum of errors in 20 trials computed separately for each auditory condition served as the basic dependent variable, assuming a minimum of zero, and a potential maximum of 380 (20 trials $\times$9 digits).

Figure 1 shows the mean number of errors obtained with the three ''irrelevant'' backgrounds. On the average, 50 errors were produced in the quiet condition, 74 while exposed to Japanese speech, and 52 with continuous pink noise, thus yielding a highly significant effect of the auditory background [$F(2,142)=97.16$; $p<0.001$]. As is evident in Fig. 1, this effect is almost entirely due to the increased error rate with speech, the two control conditions (silence and pink noise) do not produce significantly different error rates.

Presenting these overall effects for fairly standard experimental conditions only serves to make the point that the data obtained in the present investigation are entirely consistent with the literature. The effect size, an increase in error rate by about 50%, is somewhat larger than typically reported (Jones *et al.*, 1996), which may be due to the mixed presentation of auditory conditions, and to the fact that an extra retention interval delayed subject's recall, two measures, which tend to increase ''irrelevant speech'' effects. The pattern of outcomes as depicted in Fig. 1, namely the lack of impairment under continuous broadband noise, is consistent with current theoretical explanations both in terms of a ''filter'' passing speechlike information (Salamé and Baddeley, 1982), and in terms of the importance of ''changing-state'' features of the auditory background (Jones and Macken, 1993; Jones *et al.*, 1996).

B. Distribution of effect sizes

Since the focus of the present investigation is on individual differences, a fundamental question is whether sufficient individual variation is observed in the paradigm under study. That is clearly the case, as is evident in Fig. 2, which shows the distribution of individual effect sizes, operationalized as the difference in errors between the speech and quiet conditions. As would be expected for a difference between two random variables, that distribution is Gaussian, with ap-

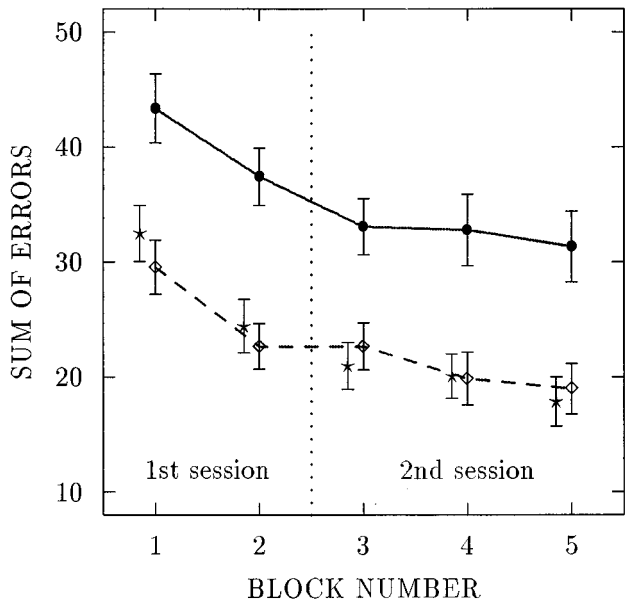

FIG. 3. Practice effects in the irrelevant speech paradigm. The graph shows the total number of errors per 30-trial block, averaged over 25 subjects, and plotted separately for the irrelevant speech condition (closed circles) and for the two control conditions: pink noise (stars), and silence (diamonds). Blocks 1 and 2 were completed in the first session, blocks 3–5 refer to data from the second session collected four weeks later.

parent deviations from a normal distribution being nonsignificant (Kolmogorov–Smirnov goodness of fit test: $z = 0.543$; n.s.).

More importantly, however, the spread of effect sizes is considerable, ranging from a minimum of -18 to a maximum of $+65$, the latter corresponding to a boost in error rate by 329% and the former reflecting an actual *reduction* in errors under speech by 33%. Note that roughly one-eighth of the sample shows no, or negative irrelevant speech effects.

Interestingly, individual effect sizes are in no way related ($r = 0.01$; n.s.) to a subject's memory capacity, which, in order to obtain a measure independent of the magnitude of the ISE, was operationalized as the number of digits recalled in the second control condition (pink noise). Further analyses did not provide any evidence for a systematic nonlinear (e.g., U-shaped) trend as a function of memory capacity either.

III. RELIABILITY OF INDIVIDUAL ''IRRELEVANT SPEECH EFFECTS''

The individual scores described in the previous section are meaningful only if they can be reliably measured. The alternative, of course, is that the distribution depicted in Fig. 2 just captures noise in the measurement procedure, not individual differences in susceptibility to the irrelevant speech effect. In order to address this problem, the individual outcomes of the experiment were treated much like scores in a psychometric test, and conventional procedures for determining the reliability of a test were applied.

A. Retest reliability

In order to examine the temporal stability of overall irrelevant speech effects within and across sessions, mean performance is depicted as a function of time in Fig. 3. In all auditory conditions, error rate drops considerably with practice, as confirmed by a highly significant main effect of block number [$F(4,96) = 15.26$; $p < 0.001$] in a two-factor analysis of variance over the five blocks and three sound conditions. The *differences* in error rates between the sound conditions, however, remain essentially the same, as indicated by the parallel curves in Fig. 3 and by the insignificant [$F(8,129) = 0.86$; n.s.] interaction between block number and sound condition in the analysis of variance. Although subjects learn to memorize more digits, they do not improve in dealing with the ''irrelevant sound,'' a finding which is in line with two other published studies investigating habituation effects within sessions (Jones *et al.*, in press) and over a two-week interval (Hellbrück *et al.*, 1996). For subsequent analyses, it justifies the use of difference scores for measuring noise effects.

Test–retest reliability (r_{tt}) captures the stability of *individual* test scores over time, and is obtained by correlating observations made on a set of subjects on two occasions (cf. Kline, 1993). Since the focus of the present investigation is on the reliability of noise effects, the *difference* in errors between the silence and speech conditions obtained for each of the 25 subjects participating in the retest was correlated with the corresponding difference obtained four weeks later. The test–retest correlation was significant ($r_{tt} = 0.45$; $p < 0.05$) but only of moderate magnitude. Interestingly, retest reliability of the ''pink-noise silence'' difference was essentially zero ($r_{tt} = -0.09$; n.s.), suggesting that performance rankings in the pink-noise control condition vary randomly over time and do not characterize individuals, a finding which agrees with earlier research employing broadband noise (e.g., Smith *et al.*, 1981).

To conclude, it turns out that the individual differences measured in the irrelevant speech paradigm are replicable over a four week interval. The moderate size of the test–retest correlation may either be due to an actual temporal instability of the attribute measured, or to a low internal consistency of the test. That possibility shall be considered in the next section.

B. Internal consistency

A method of determining the precision of measurement without having to rely on temporal stability is to compute the internal consistency of a test (see Kline, 1993). The α-coefficient (Cronbach, 1951) indicates to what extent the items of a test measure the same variable.

For the purpose of measuring noise effects, an ''item'' was defined as the difference in errors between two temporally adjacent speech and quiet trials. Thus item scores ranged from -9 (no errors under speech, all nine digits wrong in silence) to $+9$ (all false with speech, no errors in silence). Twenty such item scores were obtained for the 60-trial session completed by all 72 subjects (disregarding the 20 pink-noise trials). Internal consistency turned out to be $\alpha = 0.55$. Naturally, one would expect it to be high, since all trials were generated by the same scheme, but individual trial pairs will exhibit strong random variations, thus attenuating the consistency coefficient.

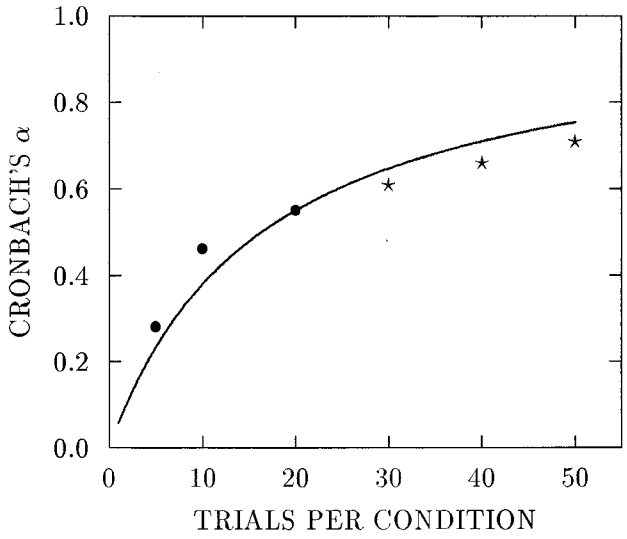

FIG. 4. Reliability of measuring the ''irrelevant speech effect'' as a function of the number of trials employed in each condition. Cronbach's α is a measure of the ''internal consistency'' of test tems (here: errors on a speech trials minus errors on a silent trial; see text). Points beyond 20 trials (stars) are based on data from those 25 subjects who participated in the retest; all other points (filled circles) are based on 72 subjects. The solid line is the improvement in ''internal consistency'' to be expected on the basis of the Spearman–Brown formula; the prediction was made based on the α of 0.55 through which the function passes.

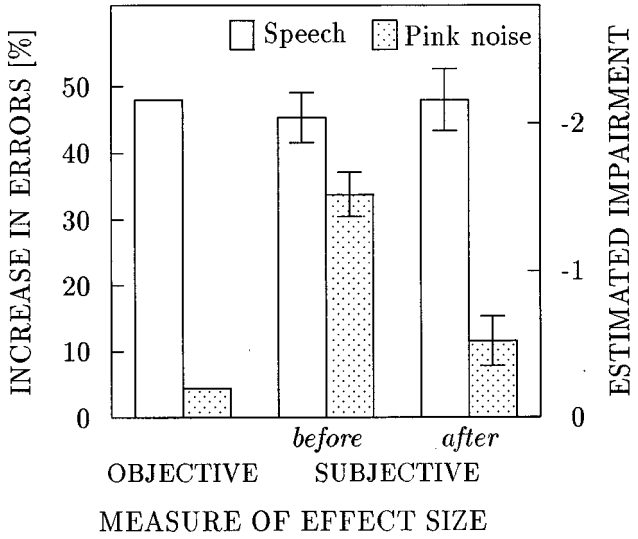

FIG. 5. Comparison of subjective estimates of memory impairment (right axis) under speech and noise (shaded bars) with objective performance decrements in percent (left axis, referring to the left pair of bars). Subjects judged expected (or experienced) effects of the sounds on a scale from -3 to $+3$ before and after performing a serial recall experiment.

Figure 4 shows how internal consistency grows with the number of speech-silence trial pairs presented to subjects. The solid line represents the theoretical prediction based on the Spearman–Brown formula describing the relationship between test length and reliability (see Nunnally, 1978, Eq. 7-7), and the data points approximate that prediction quite well. Using the Spearman–Brown formula, and extrapolating from the current α of 0.55 (which is our estimate based on the highest number of subjects and trials), we find that 66 trials per condition are needed to arrive at a reliability of 0.80, and 148 trials to reach the reliability of 0.90 considered desirable for intelligence tests, for example.

C. Temporal stability reconsidered

It turns out that much of the seemingly low test–retest reliability observed in the irrelevant speech paradigm may not be due to a temporal instability of the attribute being measured, but rather to the low internal consistency of the speech-silence error differences accumulated in a subject's score. If that is the case, one may try to estimate the underlying temporal stability of the trait (here, noise susceptibility in an ISE experiment) by correcting for the low internal consistency of the measurement procedure used. The resulting *stability coefficient* (Cureton, 1971; Lienert and Raatz, 1994, Eq. 10.10) is

$$r_{tt}(\text{stab}) = r_{tt(\text{retest})}/\alpha = 0.45/0.55 \approx 0.82, \quad (1)$$

which turns out to be fairly high, and much more encouraging than what the (uncorrected) retest coefficient suggests.

IV. SUBJECTIVE ESTIMATES OF NOISE SUSCEPTIBILITY

To address the question of whether subjects can accurately judge the performance decrements they produce in an irrelevant speech experiment, they were presented with the speech and noise samples both before and after actually performing the memory task and were asked to rate the degree to which they thought they would be affected (resp. had been affected) by the sounds.[1]

Figure 5 contrasts the objective effects of the two types of auditory materials with the mean subjective estimates of effect sizes given on two occasions. It is striking that before participating in the main experiment, subjects erroneously expect to be almost equally impaired by the speech and noise backgrounds (see the two middle bars in Fig. 5), whereas after experiencing 60 trials, the pattern of retrospectively estimated effect sizes comes much closer to the objective overall performance profile (depicted on the left in Fig. 5). This shift in the ratings shows up as a statistically significant (sound×time of testing) interaction [$F(1,24) = 20.54$; $p < 0.001$] in the two-way analysis of variance of the subjective estimates. Thus it seems that subjects can quite accurately estimate mean sound effects after participating in the experiment, while working on the assumption that ''any sound will hurt'' before.

In order to assess how well they predict their own *individual* performance changes due to noise, each subject's difference in ratings of the two sounds was correlated with the actual difference in errors produced when exposed to the sounds. While the correlation obtained with the *a priori* ratings was nonsignificant ($r = 0.16$), it increased to $r = 0.44$ ($p < 0.05$) after the experiment,[2] thus paralleling the pattern obtained for predicted mean effects.

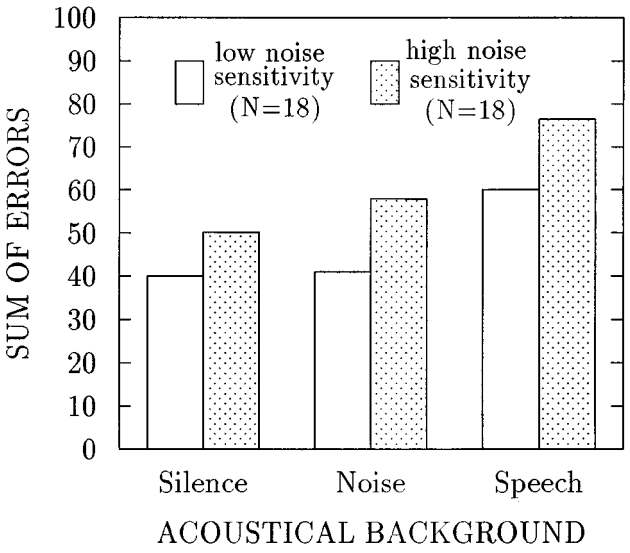

FIG. 6. Serial recall errors of subjects scoring high (upper 25%) and of subjects scoring low (lower 25%) on the noise-sensitivity questionnaire as a function of the auditory condition.

V. RELATIONSHIP BETWEEN GENERAL NOISE SENSITIVITY AND THE ''IRRELEVANT SPEECH EFFECT''

Since all subjects completed a 52-item questionnaire consisting of statements about noise in a variety of contexts (Zimmer and Ellermeier, submitted), we were able to explore whether general noise sensitivity (Weinstein, 1978) might account for some portion of the variance in the objectively measured noise effects observed in the ''irrelevant speech'' paradigm.

For that purpose, the subjects representing the highest and lowest quartile of our sample regarding their total noise sensitivity score were contrasted with respect to the errors made in the ''irrelevant speech'' task. It turned out that the highly noise-sensitive subjects produced more errors than the insensitive ones across all experimental conditions (see Fig. 6), as confirmed by the significant main effect of noise sensitivity [$F(1,34)=7.24$; $p<0.05$] in a 2×3 (sensitivity groups$\times$sound conditions) mixed analysis of variance. Planned comparisons indicate that this group difference is statistically significant ($p<0.05$) in the two sound conditions (speech and pink noise), and not significant in silence. The pattern of outcomes depicted in Fig. 6 seems to indicate, however, that noise sensitivity does not affect the error rates in the three experimental conditions differentially, which is reflected in the lack of a significant interaction in the analysis of variance.

The association between noise sensitivity and noise effects is weaker, though, than the comparison of extreme groups suggests. That becomes evident when individual noise-sensitivity scores are correlated with individual ''irrelevant speech'' effects (a given subject's difference in errors between the speech and quiet conditions). That correlation is only $r=0.23$ (significant at $p<0.05$, one-tailed test), even after correcting for the ISE's low reliability.

Interestingly, the correlation observed in the 41 female subjects is much higher ($r=0.39$) than the correlation found for the 31 male subjects in our sample ($r=0.038$), and that finding holds up when a German version of Weinstein's (1978) noise-sensitivity scale is substituted for our newly constructed questionnaire. Furthermore, the relationship between the specific impairment ratings discussed in Sec. IV and actual irrelevant speech effects also turned out to be much higher in females ($r=0.46$) than in males ($r=0.08$).

It seems that women are more accurate at judging their own noise sensitivity than are men, at least with regard to the specific effects measured in the present irrelevant speech paradigm. That is true in the absence of any overall gender effects: neither in general noise sensitivity as measured by our questionnaire do women differ significantly from men, nor in the magnitude of the ''irrelevant speech effect'' (which is 25.4 for females, 22.4 for males).

VI. DISCUSSION

The results of the present study shall be discussed with respect to three related topics: (a) the nature of individual differences in susceptibility to the ''irrelevant speech effect''; (b) the potential role of noise sensitivity in accounting for some portion of these individual differences; and (c) subject's accuracy in judging their own susceptibility to noise.

A. Nature of individual differences in ISEs

The present study establishes that—embedded in the solid overall effects typically found in ''irrelevant speech'' experiments—there are sizeable individual differences in the magnitude of the effect, which are normally distributed, spanning a range from negative effects (with facilitation due to the speech background) over null results to considerable impairment. Note that whereas in many published studies such differences might be attributable to procedural artefacts (such as the subjects receiving different orders of the ''treatments'' in a counterbalancing scheme), the present study attempted to minimize potential interactions with practice or fatigue by running all conditions randomly mixed within each block of trials.

What then, are the sources of the variance observed? Basically, we will have to consider (1) ''true'' individual differences in susceptibility to the effect, (2) measurement error, and (3) temporal instability of the variable of interest. The psychometric indices of test–retest reliability and internal consistency derived from the present data set permit to estimate the contributions of these factors to some extent.

First of all, the fact that a significant test–retest correlation (r_{tt}) was obtained suggests that there is some basic stability of the pattern of outcomes overtime. The moderate correlation of $r_{tt}=0.45$ contrasts sharply with the retest coefficients near 0.05 that Smith *et al.* (1981) obtained for a free-recall task, and with similar insignificant measures reported by Wilkinson (1974), which had suggested a haphazard rank ordering of subjects with respect to effect sizes that cannot be replicated on a second occasion. It appears that the task and sound parameters used in the irrelevant speech paradigm make it more suitable not only for showing overall effects, but also for studying individual differences in response to noise.

The internal consistency coefficient, α, on the other hand, is a measure of reliability unconfounded by temporal changes. It reflects the degree to which individual differences are captured in the same way by the items making up the test. Its magnitude, $\alpha = 0.55$, gives a better estimate of the amount of error still present in the data.

By psychometric standards, both retest reliability and internal consistency of the present measurements are disappointingly low. It turns out that this is largely due to the use of (speech-minus-silence) *difference* scores, which on statistical grounds are expected to yield lower correlations than the raw scores they are derived from whenever these raw scores are correlated themselves (see Nunnally, 1978, pp. 246–255; Lienert and Raatz, 1994, pp. 214–218). The following formula (adapted from Eq. 10.53 in Lienert and Raatz, 1994) predicts the reliability of a difference score r_{diff} from the reliabilities of the original scores (r_1, r_2) and their correlation (r_{12}):

$$r_{\text{diff}} = \frac{r_1 + r_2 - 2r_{12}}{2(1 - r_{12})}. \tag{2}$$

Given that in the present experiment recall scores under speech and in silence correlated with $r_{12} = 0.69$ and substituting the consistency coefficients for measuring errors in speech ($\alpha = 0.84$) and errors in silence ($\alpha = 0.85$) for r_1 and r_2, respectively, then the reliability of speech-silence difference scores is predicted to be $r_{\text{diff}} = 0.50$ which is quite close to the value actually obtained (0.55).

The statistical fact that the reliability of difference scores is inversely related to the correlation between the original scores creates a problem for the measurement of noise effects in terms of performance differences between experimental conditions, since that correlation is bound to be high in the irrelevant speech paradigm, given that identical memory tests are compared under two different acoustical backgrounds. In terms of psychometric theory, when working with ISE difference scores, we are not simply addressing the reliability of a ''test,'' but rather the reproduceability of a ''test profile'' which is expected to be attenuated considerably.

These considerations suggest that it is only the complement of the reliability ($\alpha = 0.85$) of the raw error scores, or a mere 15%, that make up the variance not accounted for. That residual error may be attributed to trial-by-trial fluctuations in memory span, attention, fatigue, and the like, and is to be expected even in highly homogeneous tasks like reacting to the onset of a tone repeatedly, or memorizing digit sequences as in the present experiment.

In the present context, however, we do not want to measure memory span (as reflected in the raw error scores) but rather noise effects (as reflected in a difference between error scores obtained in two conditions). Thus if a research problem requires the measurement of individual differences with a reliability comparable to that of established psychometric personality or performance tests ($\geqslant 0.90$), much larger numbers of trials will have to be collected from each subject than is commonly done to determine overall effects in an ''irrelevant speech'' experiment. Extrapolating from the theoretical curve depicted in Fig. 4 suggests that 148 trials per condition are required to arrive at a reliability of 0.90 which, given that our present 30-trial blocks took approximately 15 min, would add up to almost 3-h running time during which 148 speech and 148 silent trials would have to be intermixed. In practice these numbers may underestimate the number of trials required, as the data points falling short of the curve suggest; on the other hand, the absence of a discontinuity in the reliability estimates made after 20 and 30 trials (see Fig. 4) suggests that it is possible to pool data from sessions widely spaced in time.

B. The role of noise sensitivity

A secondary goal of the present investigation was to explore whether a person's noise sensitivity as measured by a questionnaire might account for some portion of the variance in objectively measured noise effects. While a comparison of extreme groups (Fig. 6) suggested such an influence, the overall correlation between individual sensitivity scores and ISEs turned out to be rather low ($r = 0.23$). Note, however, that noise sensitivity is a very general construct, reflecting many facets of the noise response unrelated to performance, such as sleep disturbances, interference with leisure activities, etc. On the other hand, the irrelevant speech effect very specifically measures the impact of speechlike sounds on the recall of serial order information. Thus relating a very broadly defined personality variable to a fairly narrow behavioral measure, one should not expect the relationship to be very strong. This interpretation is supported by the observation that if only those items judged *a priori* to relate to performance effects of noise are included in the correlation, it slightly increases to $r = 0.31$.

It seems that, generally, attempts to relate self-report measures of noise susceptibility to behavior have not been all that successful. Thomas and Jones (1982) found equally low correlations (averaging 0.25 across different experimental conditions) when relating noise annoyance as measured by a questionnaire to the determination of uncomfortable loudness levels in the laboratory, two measures, for which one might expect a much closer intrinsic linkeage.

As far as noise sensitivity is concerned, it might prove more promising to explore its relationship to a whole range of objective measures, using a multivariate approach more akin to the broad definition of the concept.

C. Subjective assessment of noise effects

Although self-reported general noise sensitivity did not correlate highly with the actual noise effects, one might expect a closer relationship, if subjects are queried about the specific interactions between task and noise in the experiment proper. It turns out that the subjects are unable to estimate effect sizes on the basis of familiarity with the sounds alone. They improve somewhat after gaining experience with the task. Interestingly, though, they are accurate only in predicting overall effects (see Fig. 5), while failing to predict their own noise susceptibility (as indicated by the low correlation of $r = 0.29$ between individual estimates and error rates).

Admittedly, this result might be highly dependent on the specific sounds used in a given experiment. Had we, for example, used a familiar versus an unfamiliar language as irrelevant background conditions, subjects might have predicted differential effects, while the literature suggests equal disruption (e.g., Colle and Welsh, 1976; Salamé and Baddeley, 1982). Nevertheless, a study from our laboratory (Wolski, 1996) using quite different and nonintuitive stimuli (10 varieties of frequency-modulated tones) found a similarly low correlation between estimated and observed effects ($r = 0.23$). The low validity of self-evaluations seems to be a rather general finding: Mabe and West (1982) found a mean correlation of $r = 0.29$ in their meta-analysis of 55 studies relating subject's self-evaluations to objective criteria in a number of different performance domains.

A puzzling observation contributed by the present investigation is that, based on three types of performance estimates (retrospective impairment ratings and two varieties of noise-sensitivity questionnaires), women seem to be far better at predicting their performance under noise than are men. This gender difference might deserve further systematic study.

D. Practical recommendations for measuring noise effects

The present study offers several recommendations of practical importance for the investigation of individual differences in the irrelevant speech paradigm: First of all, a considerably larger number of trials than is typically used in experiments aiming at overall effects is required: 30 trials per condition might be sufficient for measuring individual differences in error rate; several hundred trials should be collected, when differences between errors in quiet and errors under irrelevant sound constitute the variable of interest. Second, conditions should be mixed in order to avoid confounding practice and fatigue effects with the individual outcomes. Third, the difficulty of individual trials should neither be too high (ten digits to memorize) nor too low (six digits), so floor and ceiling effects do not restrict the range of individual outcomes, and thereby attenuate the internal consistency of the task. Finally, as Fig. 3 indicates, spreading data collection over several session (or weeks) does not seem to introduce discontinuities, and might be more appropriate than running long sessions incurring additional problems of attention and fatigue.

VII. CONCLUSIONS

The present analysis of objectively measured noise effects suggests the following conclusions:

(1) Invidual differences in noise susceptibility as measured in the irrelevant speech paradigm exist, and are normally distributed over a considerable range.
(2) They may be measured reliably, are fairly replicable over time, and do not change even with extensive practice.
(3) For statistical reasons, however, measurement in terms of noise-minus-quiet performance differences severely constrains the precision with which individual effect sizes may be determined.
(4) Subjects scoring high on a noise-sensitivity questionnaire produce more errors under noise than do subjects of low noise sensitivity. Nevertheless, individual differences in noise sensitivity only account for a small portion of the variance in objectively measured noise effects.
(5) Subjective estimates of the impairment produced by a specific noise source are of low criterion validity and are practically useless when subjects did not have a chance to perform under the noise in question.

ACKNOWLEDGMENTS

We would like to thank Maria Klatte for providing us with a program to run the ''irrelevant speech'' experiment using a PC sound card, and Peter Daniel from Neutrik Cortex Instruments, Regensburg, for helping with equipment calibration and loudness computations. We are grateful to Dylan Jones for his helpful comments on an earlier version of this paper.

[1]Obtaining subjective effect size ratings owes much to discussions with Dylan Jones at the occasion of the 7th Oldenburg symposium on psychoacoustics held in August, 1996.

[2]These correlations—like the ones computed for the questionnaire data presented in the next section—were corrected for measurement error due to the low reliability of the noise effects (see Sec. III B). This ''correction for attenuation'' (see Nunnally, 1978, p. 237) was applied according to the formula $\hat{r}_{12} = r_{12} / \sqrt{r_{22}}$ (see Lienert and Raatz, 1994, Eq. 11.24), where r_{12} is the ''raw'' correlation between the two variables, and r_{22} is the reliability of the second variable (here 0.55).

Buchner, A., Irmen, L., and Erdfelder, E. (**1996**). ''On the irrelevance of semantic information for the ''irrelevant speech'' effect,'' Q. J. Exp. Psychol. **49A**, 765–779.

Colle, H. A., and Welsh, A. (**1976**). ''Acoustic masking in primary memory,'' J. Verb. Learn. Verb. Behav. **15**, 17–32.

Cronbach, L. J. (**1951**). ''Coefficient alpha and the internal structure of tests,'' Psychometrika **16**, 297–334.

Cureton, E. E. (**1971**). ''The stability coefficient,'' Educat. Psycholog. Measurement **31**, 45–55.

Ellermeier, W., and Hellbrück, J. (**in press**). ''Is level irrelevant in ''irrelevant speech''? Effects of loudness, signal-to-noise ratio, and binaural unmasking,'' J. Exp. Psychol.: Human Percept. Perform.

Green, D. M., and Fidell, S. (**1991**). ''Variability in the criterion for reporting annoyance in community noise surveys,'' J. Acoust. Soc. Am. **89**, 234–243.

Hellbrück, J., Namba, S., and Kuwano, S. (**1996**). ''Irrelevant background speech and human performance: Is there long-term habituation?'' J. Acoust. Soc. Jpn. **17**, 239–247.

Job, R. F. S. (**1988**). ''Community response to noise: A review of factors influencing the relationship between noise exposure and reaction,'' J. Acoust. Soc. Am. **83**, 991–1001.

Jones, D. M. (**1993**). ''Objects, streams, and threads of auditory attention,'' in *Attention: Selection, Awareness, and Control. A Tribute to Donald Broadbent*, edited by A. Baddeley and L. Weisskrantz (Clarendon, Oxford).

Jones, D. M., and Davies, D. R. (**1984**). ''Individual and group differences in the response to noise,'' in *Noise and Society*, edited by D. M. Jones and A. J. Chapman (Wiley, New York), pp. 125–153.

Jones, D. M., and Macken, W. J. (**1993**). ''Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in short-term memory,'' J. Exp. Psychol.: Learn. Memory Cogn. **19**, 369–381.

Jones, D. M., and Macken, W. J. (**1995a**). ''Auditory babble and cognitive efficiency: The role of number of voices and their location,'' J. Exp. Psychol. Appl. **1**, 216–226.
Jones, D. M., and Macken, W. J. (**1995b**). ''Organizational factors in the effect of irrelevant speech: The role of spatial location and timing,'' Memory Cognit. **23**, 192–200.
Jones, D. M., Bridges, A., Alford, D., Macken, W. J., and Tremblay, S. (**submitted**). ''Mechanisms of auditory attention: the role of distinctiveness in the irrelevant speech effect,'' submitted to J. Exp. Psychol.: Learning, Memory, and Cognition.
Jones, D. M., and Morris, N. (**1992**). ''Irrelevant speech and cognition,'' in *Handbook of Human Performance*, edited by D. M. Jones and A. P. Smith (Academic, London), Vol. 1, pp. 29–53.
Jones, D. M., Beaman, P., and Macken, W. J. (**1996**). ''The object-oriented episodic record model,'' in *Models of Short-Term Memory*, edited by S. E. Gathercole (Psychology Press, Hove, UK), pp. 209–237.
Jones, D. M., Macken, W. J., and Mosdell (**in press**). ''The role of habituation in the disruption of recall performance by irrelevant sound,'' Br. J. Psychol.
Kline, P. (**1993**). *The Handbook of Psychological Testing* (Routledge, London).
LeCompte, D. C. (**1994**). ''Extending the irrelevant speech effect beyond serial recall,'' J. Exp. Psychol.: Learn. Memory, Cogn. **20**, 1396–1408.
Lienert, G. A., and Raatz, U. (**1994**). *Testaufbau und Testanalyse* [Test construction and test evaluation] 5th ed. (Beltz, Weinheim, Germany).
Mabe, P. A., and West, S. W. (**1982**). ''Validity of self-evaluation of ability: Review and meta-analysis,'' J. Appl. Psychol. **67**, 280–296.
Nunnally, J. C. (**1978**). *Psychometric Theory* (McGraw-Hill, New York), 2nd ed.
Salamé, P., and Baddeley, A. D. (**1982**). ''Disruption of short-term memory by unattended speech: Implications for the structure of working memory,'' J. Verb. Learn. Verb. Behav. **21**, 150–164.
Smith, A. P., and Jones, D. M. (**1992**). ''Noise and performance,'' in Handbook of Human Performance, edited by D. M. Jones and A. P. Smith (Academic, London), Vol. 1, pp. 1–28.
Smith, A. P., Jones, D. M., and Broadbent, D. E. (**1981**). ''The effects of noise on recall of categorized lists,'' Br. J. Psychol. **72**, 299–316.
Staples, S. L. (**1996**). ''Human response to environmental noise: Psychological research and public policy,'' Am. Psycholog. **51**, 143–150.
Taylor, S. M. (**1984**). ''A path model of aircraft noise annoyance,'' J. Sound Vib. **96**, 243–260.
Thomas, J. R., and Jones, D. M. (**1982**). ''Individual differences in noise annoyance and the uncomfortable loudness level,'' J. Sound Vib. **82**, 289–304.
Weinstein, N. D. (**1978**). ''Individual differences in reactions to noise: A longitudinal study in a college dormitory,'' J. Appl. Psychol. **63**, 458–466.
Wilkinson, R. T. (**1974**). ''Individual differences in response to the environment,'' Ergonomics **17**, 745–756.
Wolski, U. (**1996**). ''Experimentelle Untersuchung der Wirkung frequenzmodulierten Hintergrundschalls auf die Leistung in einer Gedächtnisaufgabe'' [''Effects of frequency-modulated tones on performance in a memory task.''] Master's thesis, University of Regensburg, Germany.
Zimmer, K., and Ellermeier, W. (**submitted**). ''Konstruktion und Evaluation eines Fragebogens zur Erfassung der individuellen Lärmempfindlichkeit'' [''Construction and evaluation of a noise sensitivity questionnaire.''], submitted to Diagnostica.