

On testing for stochastic dissociations

RUSSELL A. POLDRACK
Stanford University, Stanford, California

Methods for examining stochastic relationships have been proposed as powerful ways to dissociate different underlying psychological processes, but a number of problems have undermined conclusions based upon these methods. These testing methods and criticisms are reviewed, and the statistical methods for the testing of stochastic dependence are examined using computer simulation. With each of the methods examined, there were difficulties dealing with certain situations that are likely to occur in experiments examining dependence between successive tests. Examination also showed that the sample sizes of some previous studies were insufficient for findings of moderate amounts of dependence, calling some conclusions of stochastic independence into question. The results of the studies presented here suggest that testing for statistical dependence is a statistically perilous technique, but they also suggest several ways in which dedicated users of this form of testing can strengthen its application.

A dominant tactic in psychology is to divide the mind into a number of functional units, or modules, and then to outline relationships among those separate units (Fodor, 1983). Having proposed divisions in the cognitive architecture, one then searches for evidence that the functions of the putatively separate modules are indeed separate. This often consists of the search for *dissociations* between tests that are thought to rely primarily upon one module or the other. This approach has been used to great effect recently in the study of memory (Cohen & Eichenbaum, 1993; Roediger & McDermott, 1993), language (Caplan, 1992), and attention (Posner & Peterson, 1990), and is one of the basic tenets of the newly emerging field of cognitive neuroscience (Gazzaniga, 1994).

Dissociations between psychological tasks come in three basic forms. *Functional dissociations* are those in which two tests are differently affected by the same variable. A compelling example of functional dissociation comes from a study by Jacoby (1983). Subjects studied words in three different conditions designed to vary in the amount of conceptual or perceptual processing involved: They read the word alone, read the word in the context of an antonym, or generated the word from its antonym. Subjects were then given two different tests: a recognition test and a perceptual identification test (in which the word was presented for a very brief period to make identification difficult). Ja-

coby found that the recognition and identification tests were oppositely affected by the encoding variable, leading to the conclusion that the two tests involved different memory processes. This approach has been used extensively in recent studies of memory (see Roediger & McDermott, 1993, for review).

Tasks can also be dissociated *neurologically* by demonstrating that people with a certain form of brain damage are impaired at performing one test but perform another test just as well as normal people. (Neurological dissociations are really a subset of variable dissociations, where the variable is brain damage rather than a manipulated independent variable.) Neurological dissociations have also been examined extensively in the study of memory, and the findings often parallel those of studies using functional dissociation. For example, people with amnesia (following damage to the hippocampal system) are severely impaired relative to normal people at performance on direct memory tests such as recognition and recall, but can exhibit normal performance on indirect tests of memory such as perceptual identification (see Cohen & Eichenbaum, 1993, for review).

A third form of dissociation is the *stochastic*, or statistical, dissociation. This involves the demonstration that performance of a given person on a given item in a given task is statistically unrelated to (i.e., does not predict) performance of the same person on the same item in another task. Testing for stochastic dependence between psychological tasks is one of the most controversial methodological issues in psychology. Intuitively, stochastic independence would seem to represent a powerful way to dissociate performance on two tests; if performance for a given subject and a given item on one test is independent (i.e., not predictive) of performance of the same subject and item on another test, there is a strong presumption that these tests involve different mechanisms. Tulving (1985, p. 395) made the following claim:

Evidence provided by stochastic independence is somewhat more compelling [than evidence provided by functional

This research was supported by a predoctoral National Research Service Award from NIMH (MH10433) to the author and by a grant from NIMH (R01-MH53673) to John Gabrieli. I would like to thank Neal Cohen, Gordon Logan, and Stanley Wasserman for helpful comments on an early draft of the paper, and Arthur Flexser, John Gardiner, Arne Ostergaard, and Richard Schweickert for their helpful reviews. I would also like to thank Bill Svec for stimulating my interest in this issue and commenting on an early draft of the paper. The program and source code for the simulations presented in this article are available upon request from the author or by anonymous FTP (<ftp://golgi.stanford.edu/pub/poldrack/simulation.sea.hqx>). Correspondence should be addressed to R. A. Poldrack, Department of Psychology, Stanford University, Jordan Hall, Stanford, CA 94305 (e-mail: poldrack@psych.stanford.edu).

dissociation]: Stochastic independence cannot be explained by assuming that the two comparison tasks differ in only one or a few operating components (information, stages, processes, mechanisms). As long as there is any overlap in those operating components that are responsible for differences in what is retrieved, some positive dependence between the measures should appear. Perfect stochastic independence implies complete absence of such overlap.

This intuition has led to the use of stochastic dependence testing in a number of different debates centered around whether performance on multiple tests involves the same underlying processes. However, the use of these methods has occasioned strong critiques from those who believe that findings of independence are necessarily given to artifact (e.g., Hintzman, 1980). Hintzman (1991, p. 345) described the shortcomings of contingency analysis (i.e., testing for stochastic dependence) with the following analogy:

Memory researchers may be attracted to contingency analysis because its simplicity suggests a simple reality—in accordance with the folk psychological maxim, “Out of sight, out of mind.” But trying to measure the essential stochastic relation between two tasks with a contingency table is like trying to measure the weight of a man bearing an unknown amount of lead in his pockets and holding the tether of a helium-filled balloon. The scale may give a single, consistent reading, but that does not mean that it should be believed.

With such spirited rhetoric filling the pages of our journals, it is difficult for the dispassionate reader to evaluate claims about stochastic independence.

The aim of this article is to undertake an examination of stochastic dependence testing while attempting to avoid the heated rhetoric that has so often accompanied discourse on this issue. The primary focus of the review is on the methods used for stochastic dependence testing and some possible problems with those methods. To understand the statistical behavior of these methods, Monte Carlo simulations of a typical experiment examining stochastic relations were used. These simulations take questions about dependence testing out of the realm of theoretical statistics and demonstrate the characteristics of these methods in a way that is easy to understand. Finally, the review touches on methods that attempt to address some of the problems with stochastic dependence testing. The review will focus on the use of dependence testing in “recognition/identification” studies in memory research, but has obvious extensions to any psychological discipline in which independent processes are proposed.

What Is Stochastic Independence?

Stochastic independence between performance on two measures is a situation in which performance by a given subject on a given item in one test does not predict performance of that same subject on the same item on another test, where “performance” usually means success or failure on the task (e.g., a previously studied item is either recognized or not). This is easily understood in terms of the conditional probabilities of success or failure on each task. In an experiment examining stochastic relations, items are presented once in each of two tasks. From the data are

calculated the probabilities of success on Task 1 (P_1) and Task 2 (P_2), along with the conditional probability of success on Task 2 given success on Task 1 for each subject-item ($P_{2|1}$). If performance on the two tests is stochastically independent, then $P_{2|1} = P_2$; that is, knowing whether the subject succeeded on Task 1 does not offer any information about whether the subject is likely to succeed or fail at the same item on Task 2.

The examination of stochastic relations has been instrumental in a number of debates in the memory literature (Hintzman, 1980), but has aroused greatest interest recently with the advent of the distinction between implicit and explicit memory (for review, see Roediger & McDermott, 1993). An early example of the use of dependence testing in this arena comes from Tulving, Schacter, and Stark (1982). In their study, participants studied word lists and were then tested with both recognition and word fragment completion tests. Contingency analyses showed that, when the completion task followed the recognition task, performance on the two tests was stochastically independent; the joint probability of success on the recognition and fragment completion tasks was not significantly different from that expected under stochastic independence. Tulving et al. concluded, on the basis of these data and convergent functional dissociations, that recognition and fragment completion priming depended on different memory systems or processes.

Statistical Tests for Dependence

The measurement of stochastic relationships between psychological variables involves an examination of the relationships among the four cells of the contingency table that represents the outcomes of performance on each of the compared tests. Table 1 presents a generalized contingency table for two tests. Independence is observed when the cell probabilities are equal to those predicted by multiplying the marginal probabilities. Bishop, Fienberg, and Holland (1975) noted that nearly all measures of association (dependence) for 2×2 tables reduce either to functions of the Pearson chi-square test or to functions of the odds ratio (or cross-product ratio).

Chi-square. The most common measure for dependence in a contingency table is a version of the Pearson chi-square test (Hays, 1988), computed for a 2×2 table as:

$$\chi^2 = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (c+d) \times (a+c) \times (b+d)}, \quad (1)$$

where a , b , c , and d refer to the cell probabilities from the contingency table and N represents the number of observations.¹ The significance of this statistic is tested using a chi-square distribution with one degree of freedom; failure to exceed the critical value prevents rejection of the null hypothesis that the two variables are independent. The chi-square test for a 2×2 table is directly related to the ϕ coefficient of contingency ($\phi = \sqrt{\chi^2/N}$), and is also a direct relative of the Pearson correlation coefficient (Bishop et al., 1975).

As a test for dependence in 2×2 tables, the chi-square statistic suffers from two problems. First, it is sensitive

Table 1
2 × 2 Contingency Table, With Marginals

Task 2	Task 1		
	Success	Failure	
Success	a	b	$P_2 = a + b$
Failure	c	d	$1 - P_2$
	$P_1 = a + c$	$1 - P_1$	1.0

to the marginal proportions for each task; the test is most powerful when these marginals are equal for the two tasks, and certain configurations of marginal proportions can limit the range of the statistic (Bishop et al., 1975). In some cases sensitivity to marginal proportions may be desirable, but in the case of measuring dependence between tests, one desires a measure that can find dependence no matter how the marginal proportions are arranged. Second, the chi-square test for dependence assumes that all observations are independent; this assumption is probably violated when multiple observations come from each subject. The extent to which these violations have serious consequences for dependence testing is not clear. The simulations reported here constitute an attempt to clarify this issue.

Odds ratio. A number of popular measures of association in 2×2 tables are based on the odds ratio (α), or cross-product ratio, which is estimated as:

$$\alpha = \frac{a \times d}{b \times c}. \quad (2)$$

The odds ratio varies from zero (complete negative association) to positive infinity (complete positive association), with $\alpha = 1$ representing independence. The natural log of the odds ratio is often used instead, which varies from negative infinity to positive infinity with $\log(\alpha) = 0$ representing independence. The variance of the odds ratio for large samples is estimated as:

$$\sigma^2(\alpha) = \alpha^2 \left(\frac{1}{a} \right) + \left(\frac{1}{b} \right) + \left(\frac{1}{c} \right) + \left(\frac{1}{d} \right). \quad (3)$$

It is also possible to derive formulas for the asymptotic variance of any measure that is a monotonic function of the odds ratio (Bishop et al., 1975). Some popular measures derived from the odds ratio are Yule's Q and Goodman and Kruskal's γ (which are identical for the 2×2 table).

Measures based on the odds ratio have a feature that makes them preferable to chi-square measures: They are relatively insensitive to scaling of marginal probabilities and can attain their full range regardless of the distribution of marginal probabilities. Given that the association between two tasks is the primary factor of interest, and one would like to measure this in tests that might have widely varying levels of overall performance, the lack of sensitivity to marginals makes this class of measures preferable for the examination of stochastic dependence between tests. However, tests for dependence incorporating the variance estimator described above retain the assumption of independent observations, and thus are not

recommended in situations where individual subjects contribute multiple observations.

Methods That Incorporate Variability

The two sets of methods described above seem poorly suited for the analysis of the usual psychological experiment in which one wishes to test for dependence between two variables. Significance tests for each method assume independence between observations, an assumption that is probably violated when 1 subject provides multiple observations. In addition, these tests require the (unwarranted) assumption that the parameters of the multinomial distribution do not vary among subjects (Wickens, 1993). In fact, there are several possible loci of variation across subjects, for example, variation of marginal distributions and variation of the degree of association. Wickens (1993) showed that the usual measures of association become biased when the degree of association varies among subjects, and suggested a number of alternative tests for association when between-subjects variability exists.

t test. The simplest, yet one of the most powerful, of the alternatives that takes between-subjects variability into account is a t test on log odds-ratio values for each subject. The log odds ratio is defined as

$$\log(\alpha) = \log_e \left(\frac{A \times D}{B \times C} \right), \quad (4)$$

where A , B , C , and D are cell frequencies. The log odds ratio varies from negative infinity to positive infinity, with independence represented by the zero point. Because it is possible that b or c could take the value of zero, estimated values for the mean and variance of the log odds ratio do not exist (Agresti, 1990); an estimate of the log odds ratio that is defined for all cell frequencies can be devised by adding $1/2$ to each cell frequency:

$$\log(\alpha) = \log_e \left(\frac{(A + 0.5) \times (D + 0.5)}{(B + 0.5) \times (C + 0.5)} \right). \quad (5)$$

Wickens (1993) showed that a t test on log odds-ratio values against the null hypothesis of zero (independence) is at least as powerful as other tests (e.g., log-linear models) when between-subjects variability exists. A major advantage of this method is that normal parametric statistics (such as analysis of variance [ANOVA]), which take advantage of variation between subjects, can be used with the log odds ratio, extending questions about stochastic relations to multiple groups or conditions.

Log-linear model. The hypothesis of dependence between categorical variables can be tested using a log-linear model with subjects as a variable (a detailed discussion of log-linear models is beyond the scope of this review; see Agresti, 1990, or Wickens, 1989, for overviews). This approach is analogous to a repeated measures ANOVA, and is the standard method for examining categorical data with multiple responses per subject. Following Wickens (1989), let X and Y refer to the two measures of interest and S refer to subjects. Dependence is tested in a log-linear model in one of two ways: by test-

ing the goodness-of-fit of a model that does not contain an interaction term (referred to as [XS][YS]), or through a conditional test comparing the goodness-of-fit of the no-interaction model with a model containing an interaction term ([XY][XS][YS]). Estimated values for each model are determined with an iterative algorithm, and these values are then compared to the observed values using a likelihood ratio (G^2) statistic or another statistic derived from G^2 . Wickens (1993) examined a number of possible statistics for use with the conditional test, and found that a pseudo- F statistic (described below) was least biased by variation in the level of association across subjects.

Criticisms of Dependence Testing

Throughout its history, the use of stochastic dependence testing in psychology has been criticized on several grounds. Before further examining methods for dependence testing, I will review a number of criticisms of these methods.

Acceptance of null hypothesis. The conclusion that performance on two tests is stochastically independent (which is often the theoretically interesting outcome) has usually rested upon the failure to reject the hypothesis that the tests are independent (i.e., acceptance of the null hypothesis). Thus, it is vitally important to establish that tests for dependence have enough power to find dependence when it exists given the number of subjects and items used in the study. The power of the chi-square test is dependent on such factors as marginal probabilities (e.g., Chatterjee & Delaney, 1988), suggesting that it might be compromised in at least some cases. It is not clear that this issue has been fully addressed in the literature, and it is one of the primary concerns of the simulations to be described below. If tests for dependence have adequate power to find dependence effects of a reasonable size, this concern is mitigated.

The power of statistical tests for dependence is directly related to the number of subjects and items in a study. Examination of several studies in which stochastic dependence tests were used to examine implicit and explicit memory uncovered a wide range of such numbers between studies (Eich, 1984; Hayman & Tulving, 1989a, 1989b; Schacter, Cooper, & Delaney, 1990; Schacter, Cooper, Delaney, Peterson, & Tharan, 1991; Tulving et al., 1982; Witherspoon & Moscovitch, 1989). The number of subject items (number of subjects multiplied by number of items) in each condition for which dependence was separately tested ranged from 360 to 1,920, with a median of 1,152 subject items. As the simulations below will show, studies at the low end of this range have insufficient power to find statistical dependence when it exists, calling into question their conclusions of stochastic independence.

Simpson's paradox. Hintzman (1980, 1990; Hintzman & Hartry, 1990) has been a leading critic of the use of dependence testing in the study of memory. His criticisms are based primarily on an analysis of Simpson's paradox (Hintzman, 1980), a phenomenon that can occur

when contingency tables are collapsed over subjects, items, or both. The existence of Simpson's paradox implies that relationships that occur in the collapsed table need not necessarily be present in the individual tables, and vice versa. Hintzman and Hartry showed, for example, how different sets of items chosen from the set of word fragments used by Tulving et al. (1982) exhibited different contingency relationships, which were masked by averaging the data over items. Three covariates have been identified as possibly causing these spurious effects: subject effects, item effects, and subject-item interactions. Hintzman has argued that since one must collapse across subjects, items, or both in order to analyze contingency data, these data are simply not interpretable.

If Simpson's paradox were a serious problem in studies of stochastic relations between memory tests, one would expect that such stochastic relations would be highly variable between sets of subjects and items. However, as Martin (1981) and Gardiner (1991) have pointed out, these relationships are consistently replicable in a theoretically interpretable manner in different labs across different sets of subjects and items. This suggests that Simpson's paradox may not present a serious problem for these studies. However, statistics for contingency table analyses have often been used in direct violation of their assumptions, and the possibility for error clearly exists. In addition, systematic suppressor variables (Hintzman & Hartry, 1990) or methodological suppressors (Ostergaard, 1992; Shimamura, 1985) may result in consistently replicable yet misleading outcomes of independence.

Subject-item interaction. There has been some contention regarding the role of subject-item interactions in testing for dependence. Hintzman (1980) has described subject-item interactions as a source of spurious influence on contingency test outcomes. However, Flexser (1981) pointed out that the mathematical category of subject-item interactions includes both "spurious" effects (e.g., lapses in attention or special significance of certain items) and "true" effects (dependence or independence reflecting the underlying cognitive processes), and it is not possible to distinguish between these two sources of variation. Here, I will rely on Hintzman's usage of the term *subject-item interaction* because, within the model, subject-item interactions are introduced as a spurious source of independence. However, this should not be taken as a counterclaim to Flexser's arguments, but rather as a convenient terminological choice.

Test priming. The necessity of repeated testing with the same items in order to assess stochastic dependence introduces unique problems for the assessment of test outcomes. When items are presented twice, part of the facilitation on the second test can come from the test presentation rather than from the study presentation. For example, when a fragment completion test follows a recognition test, as used by Tulving et al. (1982), a portion of the fragment completion priming can be due to the recognition presentation. When this happens, stochastic dependence in the initial data can be masked by the test-priming effect. This is demonstrated in Table 2 (after

Table 2
Contingency Table Demonstrating Test Priming
(Adapted From Shimamura, 1985)

Task 2	Task 1					
	Pretest ($\alpha = 3.05$)		Test primed ($\alpha = 1.94$)			
	Success	Failure	Success	Failure		
Success	.34	.12	.46	.47	.26	.73
Failure	.26	.28	.54	.13	.14	.27
	.60	.40	1.00	.60	.40	1.00

Table 3
Contingency Tables for Independence
and Maximum Dependence

Task 2	Task 1					
	Independence ($\alpha = 1.0$)		Maximum Dependence ($\alpha = 3.3$)			
	Success	Failure	Success	Failure		
Success	.42	.28	.70	.48	.22	.70
Failure	.18	.12	.30	.12	.18	.30
	.60	.40	1.00	.60	.40	1.00

Note— $P_1 = .60$, $P_2 = .70$, and $P_{NS1} = .40$. The cell a probability for maximum dependence is calculated as $a = P_1 - P_{NS1} + (P_{NS1} \times P_2)$ where P_{NS1} is the probability of success for nonstudied items on Task 1 (Ostergaard, 1992). The rest of the cell probabilities were calculated according to Table 1.

Shimamura, 1985). In the initial table, there is a positive relationship between the two tasks, and the chi-square test against independence is significant. In the test-primed table, half of the previously uncompleted items on Task 2 are now completed (regardless of their recognition outcome). The amount of association in the new table is reduced from $\alpha = 3.05$ to $\alpha = 1.94$, and the chi-square test against independence now fails to reach significance. This is indeed a serious problem for successive testing paradigms, but it may be relatively simple to devise studies that are not compromised by test priming.

Range restriction. The problem of range restriction, as outlined by Ostergaard (1992), arises when changes in performance due to memory are small relative to baseline performance on a task. When this is the case, the amount of possible dependence between two tests that is due to memory is limited to the portion of performance that arises from memory; independence of the portion of performance not due to memory can mask dependence in the portion due to memory, so that one may conclude that two tasks are independent even when memory-based performance is maximally dependent. Ostergaard (1992) demonstrated a method for estimating the maximum amount of dependence that could arise between two tasks due to memory, and showed that this maximum amount of dependence can be quite small when performance on studied items is only slightly above baseline performance (performance on nonstudied items). Table 3 presents the method for estimating maximum dependence from memory for a contingency table. As an example of the effects of range restriction, Figure 1 presents a plot of the maximum amount of dependence (in terms of odds

ratio) against baseline performance for a given level of performance on studied items. As the baseline approaches performance on studied items (i.e., as the effect of study decreases), the maximum amount of dependence from memory is reduced.

Ostergaard (1992) computed maximum dependence estimates for several published studies and tested these estimates against the observed data. In each of 18 study conditions examined, the confidence interval for the joint proportion of success (i.e., cell a in Table 1) contained both that proportion expected under stochastic independence and that expected under maximum dependence.² That is, it was impossible to differentiate between cases of independence and maximum dependence. He suggested that, in addition to testing for stochastic dependence between memory tests, one should test against the null hypothesis of maximal memory dependence. If one is unable to reject the null hypothesis of maximum dependence, the conclusion of stochastic independence is seriously questionable.

Addressing Problems With Dependence Testing

The discussion of problems with dependence testing has led those who are partial to such methods to devise ways to address these criticisms. These proposals have included both methodological and statistical procedures.

Homogenization procedure. In order to correct for problems that may arise due to subject and item differences within the traditional framework of analysis, Flexser (1981) introduced a procedure that corrects for these effects in collapsed contingency tables. Called the *homogenization procedure*, this method uses the correlations between sets of subject scores and (separately) sets of item scores to estimate subject and item effects on the collapsed contingency table. These effects are then removed, resulting in the table that is estimated to occur if subject or item effects had been absent. Flexser (1981) discussed several examples in which dependence due to random subject and item variation was decreased by using

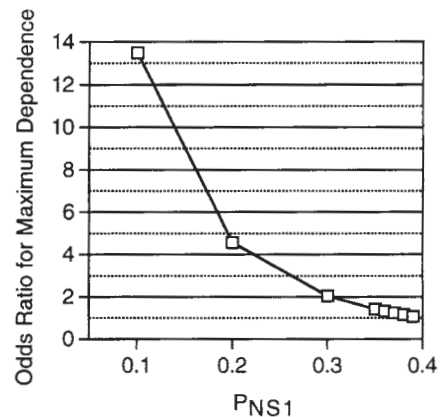


Figure 1. Odds ratio for maximum dependence as a function of the level of performance on nonstudied items (assuming $P_1 = .4$ and $P_2 = .6$).

this method. Hintzman and Hartry (1990) have claimed, however, that homogenization did not fully remove item effects in an analysis of their own contingency data. In the simulations reported below, the effects of homogenization in the presence of subject and item effects and subject-item interactions are examined.

Reducing test priming. Shimamura (1985) noted two methods that could be used to mitigate the influence of test priming on dependence analyses. First, one can examine the amount of dependence across two levels of an independent variable where test priming is equal across those levels. While this allows claims about the relative independence of two sets of processes, it does not allow the claim that two processes are independent in any absolute sense; while this may indeed be a valid conclusion (see Hayman and Tulving, 1989a, for extension of this technique), it will disappoint those who wish to show that two processes are independent.

The second way to reduce the effects of test priming is to use tests that do not allow test priming. For example, Shimamura (1985) noted that Eich (1984) used a spelling test with homophones following a verbal recognition test, which did not give any additional clues about the spelling of the tested words. With respect to studies of implicit

Table 4
List of Model Parameters That Were Specified for Each Simulation

Symbol	Parameter
N	Number of subjects and items ($N_{\text{subjects}} \times N_{\text{items}} = N_{\text{subject-items}}$)
P_1	Probability of success on Task 1
P_2	Probability of success on Task 2
α	Odds ratio
SD_{item}	Standard deviation of item variation
SD_{subject}	Standard deviation of subject variation
$P_{\text{SI}+}$	Probability of positive subject-item interaction
$P_{\text{SI}-}$	Probability of negative subject-item interaction

Note—The role of each parameter is described in the text.

and explicit memory, this strategy is enhanced by the fact that changes of modality have divergent effects on direct and indirect tests of memory; direct tests of memory such as recognition are usually affected little or not at all by changes in modality, while implicit memory is usually reduced or eliminated by changes in modality (see Roediger & McDermott, 1993). Thus, this strategy may be useful for solving the problem of test priming for studies of implicit and explicit memory.

Range restriction. The solution to the range restriction problem is decidedly simple: Use only tasks in which

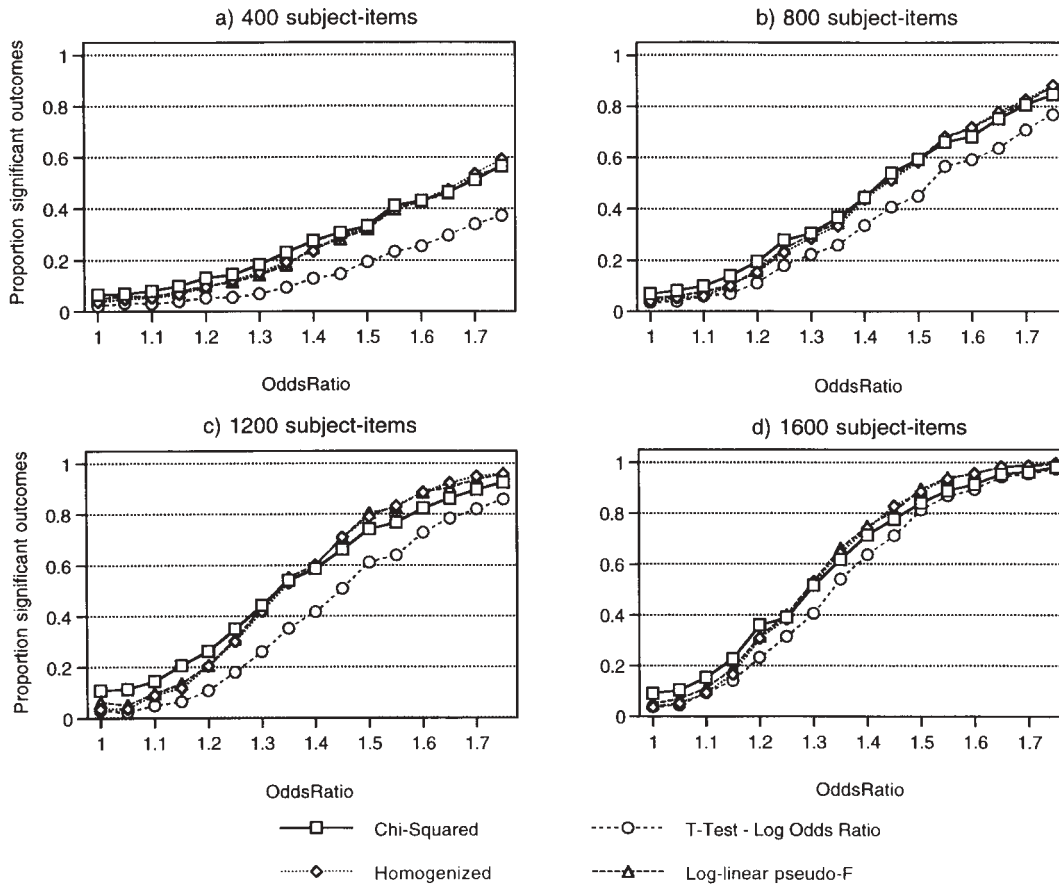


Figure 2. Proportion of significant results for each statistical test over the range of levels of dependence, from $\alpha = 1.0$ (independence) to $\alpha = 1.75$ (moderate dependence), varying the number of subject items ($P_1 = .4$, $P_2 = .6$, $SD_{\text{subject}} = 0.15$, $SD_{\text{item}} = 0.05$, $P_{\text{SI}+} = 0$, $P_{\text{SI}-} = 0$).

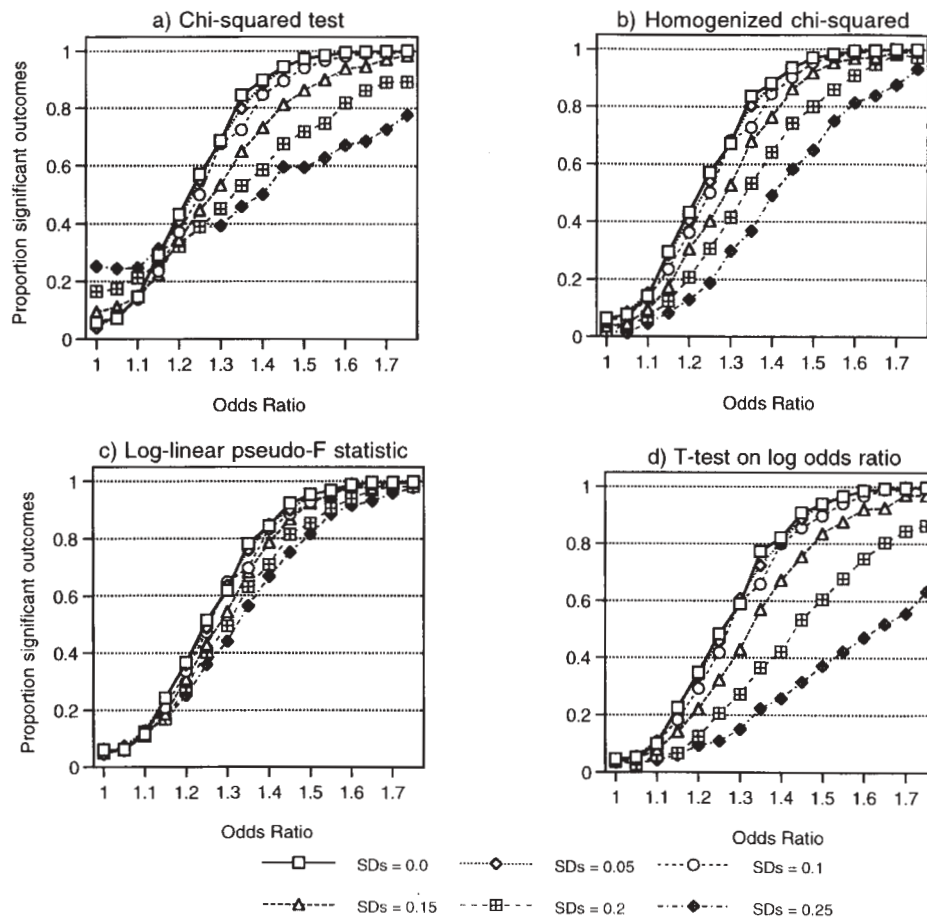


Figure 3. Proportion of significant outcomes for each statistical test as a function of the level of subject variability (SD_{subject}), varying the odds ratio ($P_1 = .4$, $P_2 = .6$, $SD_{\text{item}} = 0$, $P_{S1+} = 0$, $P_{S1-} = 0$, $N_{\text{subjects}} = 40$, $N_{\text{items}} = 40$).

study produces large improvements in performance over baseline. As Ostergaard (1992) noted, this can be difficult for some classes of tests, such as implicit memory tests, where memory effects are usually small. One approach to determining whether dependence analysis is appropriate would be to calculate the odds ratio for maximum memory dependence (as performed above; see Figure 1) and to only proceed with dependence analysis if this odds ratio exceeds some reasonable criterion. The data presented below can be used as a rough guide to setting this criterion, based on the number of observations in the study.

Simulations

In order to examine the behavior of the statistical methods that are used to test for dependence, the performance of subjects in a typical memory experiment involving successive memory tests was simulated. The goal of the simulations was to determine the degree to which the tests can find dependence when it exists (i.e., power), and the degree to which the tests claim to find dependence when it does not exist (i.e., Type I error), both as a function of

the number of subjects and items used in the experiment. In addition, the simulations permitted examination of the degree to which variability in the data affected the outcome of the statistical tests, and the way in which marginal levels of performance affected test outcomes. In order to ensure that the findings of the simulation were applicable to memory studies, the characteristics of the simulated data were based as closely as possible on studies of word fragment completion, a task that is often used in conjunction with dependence testing (e.g., Hayman & Tulving, 1989b; Tulving et al., 1982).

The simulations were performed on a Power Macintosh microcomputer. Uniform random deviates were produced with a multiplicative congruential method with a Bays-Durham shuffle (Press, Teukolsky, Vetterling, & Flannery, 1992, procedure ran1), and normal (Gaussian) deviates were produced by applying the Box-Muller method to these uniform deviates (Press et al., procedure gasdev). Each run simulated over 1,000 pseudoexperiments, with a new random seed provided for every run.

Variance parameters were set in order to equate the amount of overall variance with that reported for studies

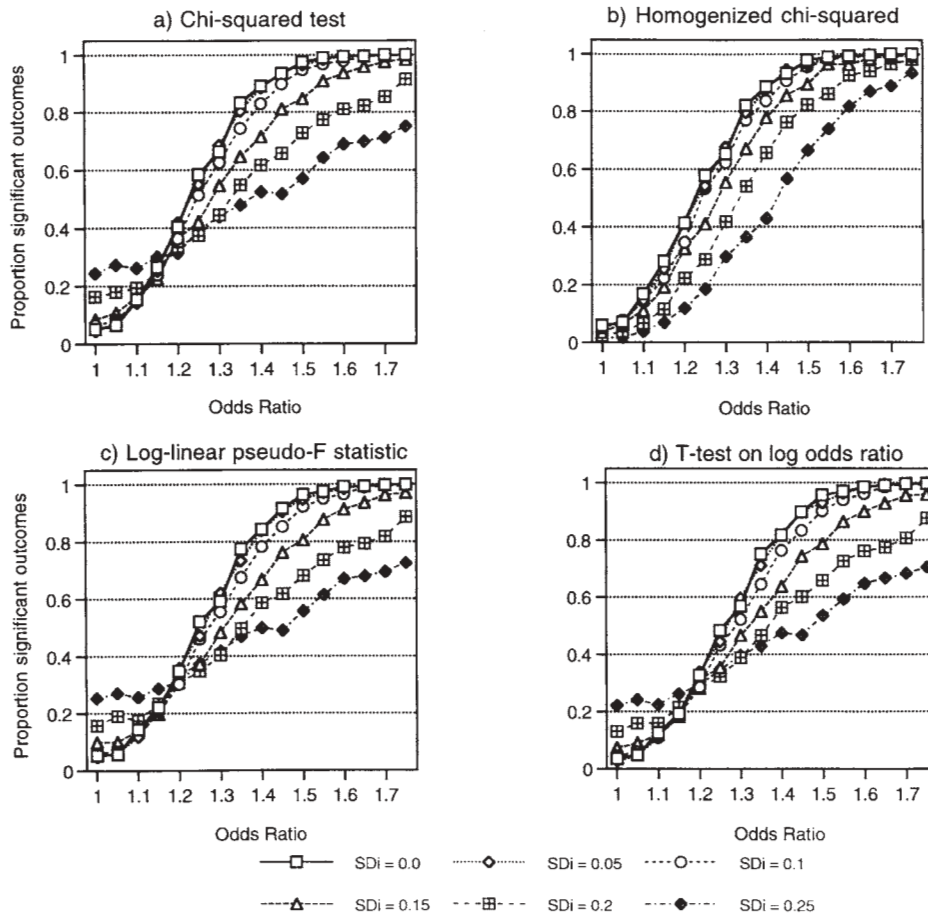


Figure 4. Proportion of significant outcomes for each statistical test as a function of the level of item variability (SD_{item}), varying the odds ratio ($P_1 = .4, P_2 = .6, SD_{subject} = 0, P_{SI+} = 0, P_{SI-} = 0, N_{subjects} = 40, N_{items} = 40$).

of word fragment completion (e.g., Sloman, Hayman, Ohta, Law, & Tulving, 1988).³ Subject-item interactions (when present) were introduced in two ways, corresponding to positive and negative interactions. For positive interactions, it was stipulated that the outcome of the two memory tests on a given trial were either both successes or both failures; this is the sort of interaction that would occur due to items with special significance or special difficulty for a given subject. For negative interactions, it was stipulated that the outcome on one test was a success and the other was a failure. A negative interaction might occur, for example, due to a lapse of attention on the part of the subject during one presentation of an item.

Table 4 presents the parameters that were provided for each simulation (these parameters will be introduced as the model is described). The algorithm used to simulate performance on each trial proceeded as follows:

1. The values of each of the parameters listed in Table 4 were specified.

2. Using normal random deviates with mean of zero and standard deviations of $SD_{subject}$ and SD_{item} , respectively, deviations were created for each subject across all items ($SubDev_i$) and for each item across all subjects

($ItemDev_j$). These deviations allowed the independent introduction of subject and item variation.

3. For each subject i and item j , marginal proportions P_{1ij} and P_{2ij} (for Tasks 1 and 2, respectively) for a given trial were determined as follows:

$$P_{1ij} = P_1 + SubDev_i + ItemDev_j$$

$$P_{2ij} = P_2 + SubDev_i + ItemDev_j$$

In the case that any marginal proportion was calculated as less than .0001 or greater than .9999, the probability was set equal to the appropriate one of those bounding values.

4. Using the marginals from Step 3 along with the selected value of the odds ratio parameter, cell probability a was calculated using Equation 7, and the rest of the cell probabilities were calculated using the standard formulas listed in Table 1. In the case that any cell probability was calculated as less than .0001 or greater than .9999, the probability was set equal to the appropriate bounding values.

5. The outcome of each simulated trial was determined by using a uniform random variate to assign the trial to one of the four cells in the contingency table, on the basis of the probabilities calculated in Step 4.

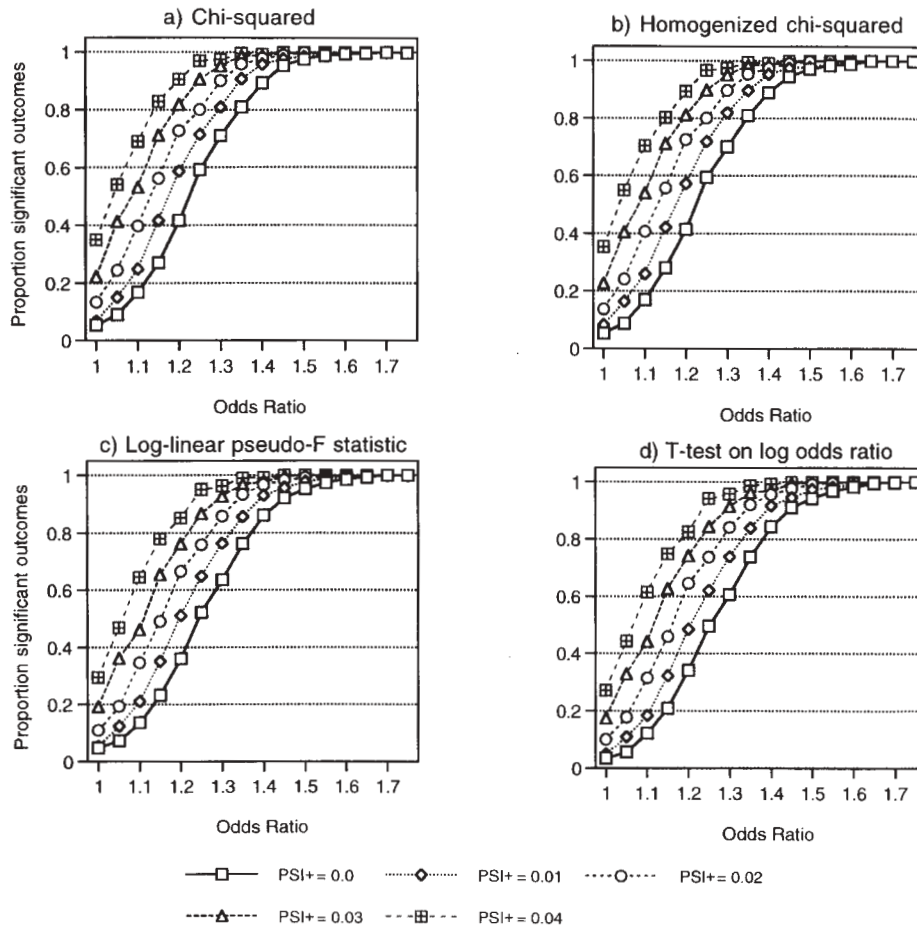


Figure 5. Proportion of significant outcomes for each statistical test as a function of the probability of positive subject-item interactions, varying the odds ratio ($P_1 = .4, P_2 = .6, SD_{\text{subject}} = 0.15, SD_{\text{item}} = 0.05, P_{SI-} = 0, N_{\text{subjects}} = 40, N_{\text{items}} = 40$).

6. If subject-item interactions were present, there was probability P_{SI+} that the outcome would be reassigned (with equal probability) to either cell *a* or cell *d* of the contingency table, and there was probability P_{SI-} that the outcome would be reassigned to either cell *b* or cell *c* of the contingency table.

After creation of the simulated data set, the data were analyzed using the procedures described above and used regularly in the literature. Overall performance (averaged over subjects) was used to compute the chi-square statistic. Flexser's (1981) homogenization method for the removal of subject and item effects was applied to the data, and the chi-square statistic was computed from the transformed table. These statistics were tested against the null hypothesis of independence using a chi-square distribution with 1 degree of freedom. A log-linear model was fitted to the data with subjects as a factor using the iterative proportional fitting algorithm (Wickens, 1989). The hypothesis of dependence was tested using a pseudo-*F* statistic based on the log-linear G^2 statistics:

$$F_g(1, n-1) = \frac{(n-1) \times (G^2_{[XS][YS]} - G^2_{[XY][XS][YS]})}{G^2_{[XY][XS][YS]}}, \quad (6)$$

where n is the number of subjects. Wickens (1993) found that this statistic was less affected than were other log-linear statistics by variation in association between subjects. The log odds ratio was computed for each subject across items, and these were tested across subjects against a mean of zero using a one-sample *t* test.

For each set of parameters, 1,000 pseudoexperiments were run, in which each pseudoexperiment consisted of one replication with number of subjects and items specified by the parameters. Each of these runs yielded 1,000 values for each statistical test, and the number of statistically significant outcomes for the run was tabulated. Estimates of the number of significant outcomes were reasonably stable across runs with this number of pseudoexperiments. The proportion of significant outcomes signifies Type I error when no dependence is present, and signifies the power to find a given level of dependence when such dependence is present.

Due to the large search space, only a subset of possible parameter combinations was examined. When parameters were not varied systematically, they were set in order to reproduce as closely as possible the data from Tulving et al. (1982) or other studies of stochastic relations between

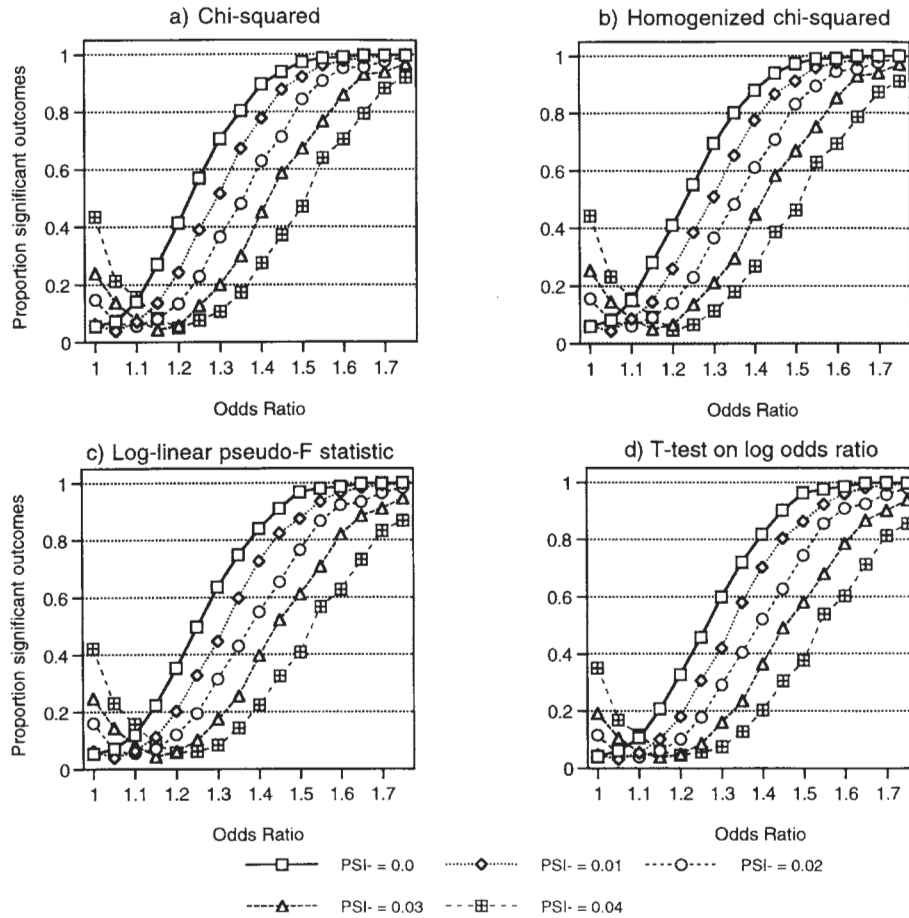


Figure 6. Proportion of significant outcomes for each statistical test as a function of the probability of negative subject-item interactions, varying the odds ratio ($P_1 = .4, P_2 = .6, SD_{subject} = 0.15, SD_{item} = 0.05, P_{SI+} = 0, N_{subjects} = 40, N_{items} = 40$).

memory tests. One parameter that deserves explanation is the odds ratio (α). This is an effect size measure for the amount of dependence in the data. This measure is computed from a contingency table by Equation 2. To go from the odds ratio to a contingency table, Equation 7 (see below) was used to determine the cell a proportion.⁴

Cell proportions for other cells were then calculated from the marginal proportions. In order to create a table for independent tasks, the odds ratio was set to 1.0; for dependence, the odds ratio was increased.

RESULTS AND DISCUSSION

Unless otherwise noted, all figures present the proportion of significant outcomes from each set of 1,000 simulations. Except for simulations in which the number of subjects and items was explicitly manipulated, they

were set to 40 subjects and 40 items per simulation (1,600 subject items), which falls within the range of subject items in the dependence studies examined above.

Error and power of tests. Using levels of subject and item variation ($SD_{item} = 0.05, SD_{subject} = 0.15$) that produced reasonable overall variance ($SD_{overall} \approx 0.15$), the significance profiles of each statistical test for dependence were simulated, varying the number of subject items; different levels of SD for subjects and items were chosen on the basis of model outcomes rather than on the basis of independent evidence, and in the studies presented below, variation in both of these parameters was examined separately. Profiles were obtained by varying the dependence parameter over a range that covered the full range of the tests (from $\alpha = 1.0$ to $\alpha = 1.75$). These significance profiles are plotted in Figure 2. The most impressive (though not surprising) finding from this analysis is the effect of

Equation 7

$$a = \frac{1 - P_1 + \alpha \times P_1 - P_2 + \alpha \times P_2 - \sqrt{-4 \times (\alpha - 1) \times \alpha \times P_1 \times P_2 + (P_1 - \alpha \times P_1 + P_2 - \alpha \times P_2 - 1)^2}}{2 \times (\alpha - 1)}$$

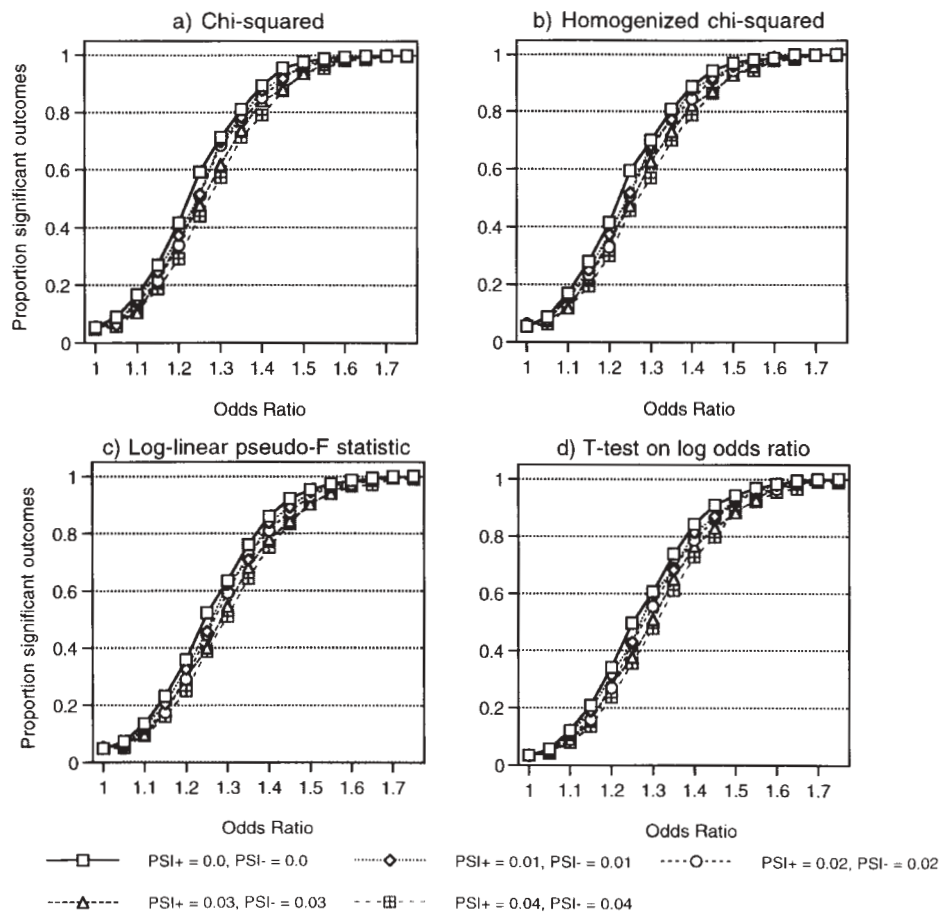


Figure 7. Proportion of significant outcomes for each statistical test as a function of the probability of positive and negative subject–item interactions, varying the odds ratio ($P_1 = .4$, $P_2 = .6$, $SD_{\text{subject}} = 0.15$, $SD_{\text{item}} = 0.05$, $N_{\text{subjects}} = 40$, $N_{\text{items}} = 40$).

the number of subject items on the tests. With 400 subject items (Figure 2a), the tests had power of less than 60% to detect even a relatively large effect ($\alpha = 1.75$). With 1,600 subject items (Figure 2d), on the other hand, each of the tests had greater than 80% power to find an effect of size $\alpha = 1.5$, and power greater than 90% to find dependence effects of size $\alpha = 1.75$. While power increased with every increase in sample size, it was most dramatically increased as the number of subject items increased from 400 to 800. This suggests that the use of subject–item sample sizes of less than 800 results in greatly reduced power to find dependence (and therefore bias towards independence). Given this result, findings of independence from studies using such small sample sizes are suspect.

In terms of the individual tests, two important trends were evident. First, as the number of observations increased, the level of Type I error for the chi-square test increased above the nominal .05 level; with 1,200 subject items, Type I error was greater than 10%. As tests below will show, this finding was directly related to variation among subjects and items. Second, the *t* test on log

odds ratios exhibited less power than the other tests as dependence increased. This occurred because of the use of a two-tailed test against the null hypothesis of $\log(\alpha) = 0$. Because the log odds ratio varies around zero with both directions representing dependence, the two-tailed test seemed most appropriate for examining the question of independence.

Effects of subject and item variability. The effects of subject and item variability on each of the statistical tests were examined by varying the levels of each while keeping other parameters constant. The results of these simulations are presented in Figures 3 and 4 (manipulating subject and item variability, respectively). Type I error increased with the amount of subject and item variation for the chi-square test, and power was decreased with greater variation. The homogenization procedure was able to correct for the effect of subject and item variability; Type I error remained at or below the nominal level across the range of subject and item variation when the data were homogenized. However, when homogenization was used with the chi-square test, increases in subject and item variability resulted in decreasing power to find dependence.

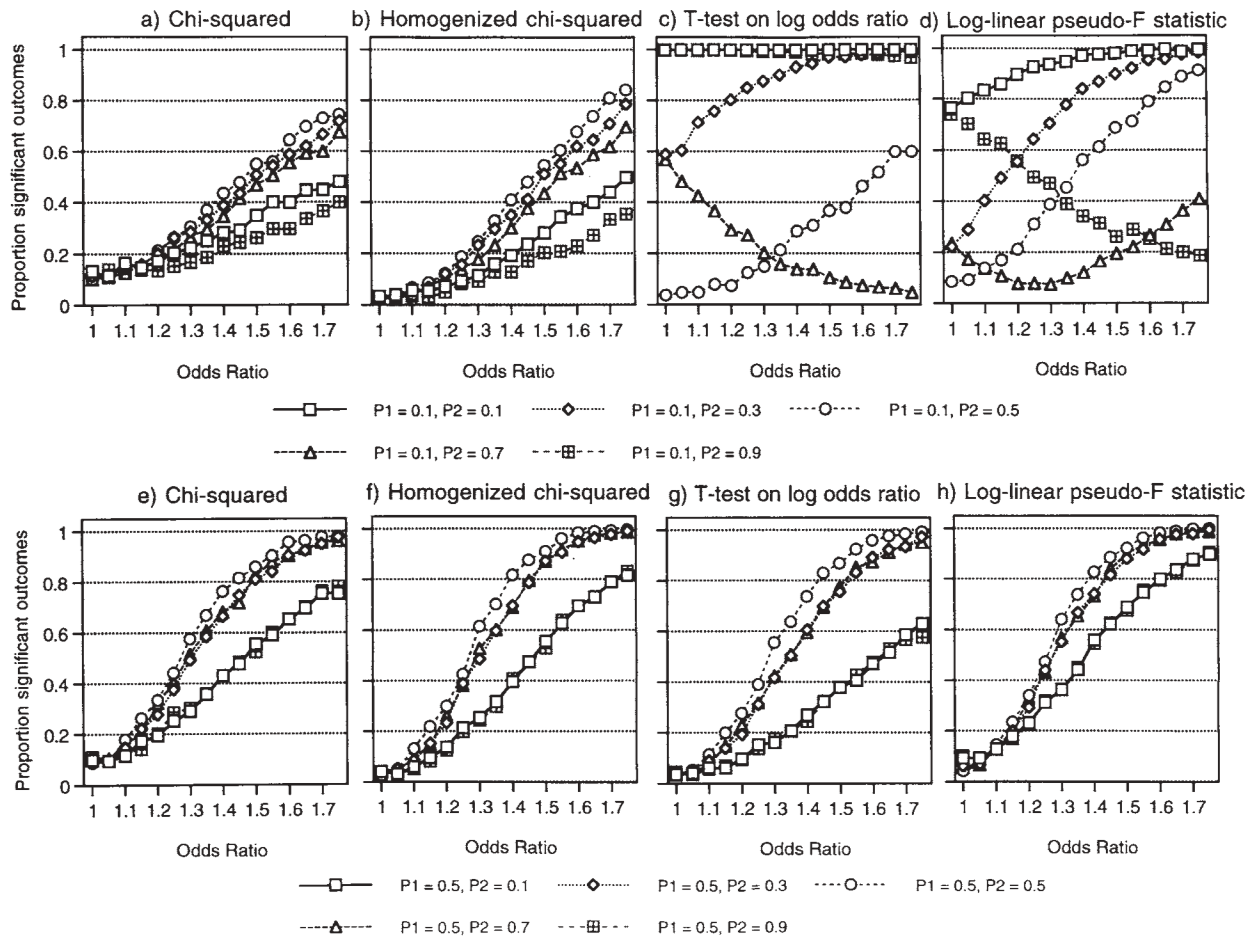


Figure 8. Proportion of significant outcomes as a function of marginal probabilities for each statistical test. Results are split into two panels, representing $P_1 = .1$ and $P_1 = .5$, respectively ($SD_{\text{subject}} = 0.15, SD_{\text{item}} = 0.05, P_{SI+} = 0, P_{SI-} = 0, N_{\text{subjects}} = 40, N_{\text{items}} = 40$).

The t test on log odds ratios was differently affected by subject and item variation. This was expected, since responses were averaged over items, and the hypothesis test was performed across subject means. Subject variation did not affect Type I error, but had a strong effect on power. Since the value of the t test varies inversely with the observed standard deviation, and this value was taken across subjects, it is natural that increases in variability should decrease power. Increases in item variation had detrimental effects on both Type I error and power, in a manner similar to the chi-square test on raw data.

The log-linear F_g statistic performed well in the face of subject variation, losing only a small amount of power with large amounts of variation. However, item variation had effects on F_g similar to its effects on the t test and chi-square test; Type I error was increased, and power was decreased, as the level of item variation increased. Thus, none of the statistics examined here performed fully well in the face of subject and item variation; the homogenization procedure with the chi-square test fared best, never exceeding the nominal Type I error level, but had decreased power with variation, as did the other tests.

Effects of subject–item interactions. The effects of subject–item interactions were examined across variations in the odds ratio. These data are plotted in Figures 5 and 6, which show the effect of increasing P_{SI+} and P_{SI-} , respectively, plotted against the odds ratio; Figure 7 presents the effect of increasing P_{SI+} and P_{SI-} together. As is evident from Figure 7, even relatively infrequent positive subject–item interactions led to significant dependence outcomes. When such interactions occurred on 4% of all trials when performance was otherwise generated by independent processes, dependence was detected by all of the tests more than 25% of the time. Homogenization of the contingency tables did not mitigate the effect of these interactions.

Negative subject–item interactions shifted the significance profile to the right for all tests. That is, in order to find independence in the presence of negative subject–item interactions, the processes generating performance must show dependence. This outcome poses a problem for the interpretation of nonsignificant test outcomes, because it suggests that relatively infrequent negative interactions can mask significant amounts of dependence be-

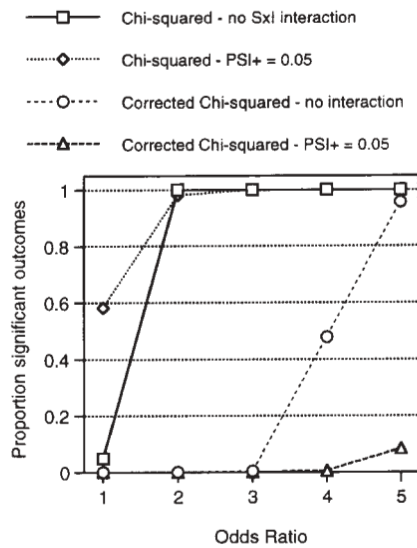


Figure 9. Proportion of significant outcomes for each chi-square test (using $N_{\text{subject-items}}$ or N_{subjects} in the numerator) as a function of the probability of positive subject-item interactions, varying the odds ratio ($P_1 = .4$, $P_2 = .6$, $SD_{\text{subject}} = 0.15$, $SD_{\text{item}} = 0.05$, $P_{\text{SI-}} = 0$, $N_{\text{subjects}} = 40$, $N_{\text{items}} = 40$).

tween underlying processes. When positive and negative subject-item interactions were present, test outcomes were relatively unaffected; Type I error was unaffected, but power was slightly decreased.

Effects of marginal probabilities. The relationship between power and marginal probabilities was examined by varying P_2 while holding P_1 constant, over a range of α values. These data are presented for each test in Figure 8. Chi-square tests on raw and homogenized data were similarly affected by marginal probabilities. Power was decreased as the marginal probabilities differed from one another, and was increased as marginal probabilities approached .5.

The t test on log odds-ratio values was more drastically affected by marginal frequencies. When either of the marginal probabilities was near .5 (as in Figure 8c), Type I error was not increased by varying marginal probabilities, but power was decreased as marginal probabilities differed and was greatest when both marginals neared .5. When both of the marginal probabilities approached the limiting values of 1.0 or .0, the results of the parametric test were erratic; for example, when both marginal probabilities were .1, the test exhibited 100% Type I errors. When the two marginal probabilities differed greatly (i.e., one was .1 and one was .9), power actually decreased as the odds ratio increased. This erratic behavior disappeared when subject and item variation in the model were turned off; it was likely due to the occurrence of marginal probabilities that were very close to the bounds of zero and one (marginal values were bounded at .0001 and .9999 in the simulation), but its source remains a puzzle. The log-linear F_g statistic behaved in a manner quite similar to the t test on log odds ratios.

Correcting the chi-square test. Ostergaard (1992) suggested that because individual observations within a subject are correlated, one should correct the chi-square test by using the number of subjects, rather than the number of subjects multiplied by the number of items, to determine the number of observations for the test. This is equivalent to analyzing subject means in the chi-square test as single data points. This correction was examined in the face of subject-item interactions; these data are presented in Figure 9. The correction was found to be overly conservative, even in the face of subject-item interactions that biased the normal chi-square test toward dependence; the power of the corrected chi-square test to find independence when it occurred was sharply attenuated, and the correction resulted in zero Type I errors. This correction should not be used, as it will result in a preponderance of Type II errors.

GENERAL DISCUSSION

In the present study, simulations were used to examine the statistical profiles of tests for stochastic dependence between memory tests. The following conclusions can be drawn from the simulations and analysis:

1. The number of subject items affected the power of each of the statistical procedures used to examine stochastic dependence. When the number of subject items was small (400), these tests had insufficient power to find dependence effects. However, when the number of subject items was large (1,600), these tests had sufficient power to find moderately large dependence effects. Thus, it is recommended that studies of stochastic dependence collect at least 1,600 subject item observations.

2. The only one of the statistics examined that did not exhibit inflated Type I error rates in the face of item variation was the chi-square test on homogenized data. For all of the tests, subject variation and item variation resulted in a decrease in power to find dependence.

3. Even infrequent subject-item interactions had strong effects on each of the tests examined, and homogenizing the data did not correct for this. Positive interactions resulted in increased Type I error, and are thus unlikely to result in spurious conclusions of independence; however, negative subject-item interactions shifted the significance profile for each test, and might mask dependence when it exists. When present together, positive and negative interactions canceled each other out with little effect on test outcomes.

4. As marginal frequencies differed from one another for two tasks, tests for dependence performed poorly. Chi-square tests (with and without homogenization) had decreased power as marginals approached the extremes of the distribution, and as marginals differed from one another. Between-subjects statistics (t test on log odds ratios and log-linear F_g statistic) showed decreased power when marginals differed, and performed erratically when marginals approached the extremes of the distribution. Care should be taken in using dependence testing when

marginals are extreme or when they differ greatly from one another.

5. The problems of range restriction (Ostergaard, 1992) and test priming (Shimamura, 1985) can seriously suppress existing dependence and thus bias measures of association toward independence, and steps should be taken to avoid them. The odds ratio for maximum memory dependence should be computed, and dependence testing should be performed only if sufficient power exists to find that level of dependence. In addition, a statistical test against maximum memory dependence should be performed whenever the test against the null hypothesis is not significant; if the test against maximum dependence is also nonsignificant, the outcome of independence is questionable.

6. The correction to the chi-square statistic, recommended by Ostergaard (1992) for use with the maximum dependence test, resulted in extreme conservatism and thus should be avoided.

Consequences of Variability

Each of the tests for dependence was affected by the presence of subject or item effects in the data. When the homogenization procedure was used with the chi-square test, these effects were minimized. If one wishes to reduce the influence of these effects, the best choice for analysis of data in successive memory test experiments is the chi-square test on homogenized data. The reservations expressed by Hintzman and Hartry (1990) regarding the ability of the homogenization procedure to correct for subject and item effects were not confirmed by the present analysis. It might be argued that this finding depends crucially on details of the simulation. However, two facts weaken this argument; the characteristics of the simulated data were modeled closely on existing data sets, and the homogenization procedure worked well across a number of manipulations that adversely affected other measures.

One surprising finding of the present study is how a small proportion of subject-item interactions can have large effects on test outcomes. When positive, these interactions bias the test in favor of dependence, and thus are unlikely to result in spurious claims of independence. Negative interactions can have two effects: they can bias the test in favor of dependence when the underlying processes are independent, and they can also bias the test against finding dependence when it really exists. The latter of these eventualities is worrisome for users of contingency analysis, because it means that conclusions of stochastic independence could be spurious. It is likely that both positive and negative interactions both occur, and the simulations suggest that they may cancel each other with little effect on overall test performance. This issue requires further analysis.

Conclusions

The results of the current study suggest that the tests used to examine dependence in 2×2 contingency tables in memory experiments may be sufficiently powerful with larger sample sizes, but they are susceptible to sev-

eral adverse effects. The effects of subject and item variability may be countered by using the method for homogenization of contingency tables suggested by Flexser (1981). Other problems, such as the influence of marginal frequencies and the problems of range restriction and test priming, may be avoided by careful choice of tasks, but other problems remain (such as the suppressive effect of subject-item interactions) that leave claims of stochastic independence on shaky ground. Although further statistical development may address some of these problems, for now the examination of stochastic relations is at best an adjunct to other techniques for establishing dissociations between mental processes.

REFERENCES

- AGRESTI, A. (1990). *Categorical data analysis*. New York: Wiley.
- BISHOP, Y. V., FIENBERG, S. E., & HOLLAND, P. W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.
- CAPLAN, D. (1992). *Language: Structure, processing, and disorders*. Cambridge, MA: MIT Press.
- CHATTERJEE, S., & DELANEY, N. J. (1988). Contingencies for analysis of contingency tables: More on the chi-square test. *British Journal of Mathematical & Statistical Psychology*, **41**, 235-249.
- COHEN, N. J., & EICHENBAUM, H. (1993). *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press.
- EICH, E. (1984). Memory for unattended events: Remembering with and without awareness. *Memory & Cognition*, **12**, 105-111.
- FLEXSER, A. J. (1981). Homogenizing the 2×2 contingency table: A method for removing dependencies due to subject and item differences. *Psychological Review*, **88**, 327-339.
- FODOR, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.
- GARDINER, J. M. (1991). Contingency relations in successive tests: Accidents do not happen. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 334-337.
- GAZZANIGA, M. (Ed.) (1994). *The cognitive neurosciences*. Cambridge, MA: MIT Press.
- HAYMAN, C. G., & TULVING, E. (1989a). Contingent dissociations between recognition and fragment completion: The method of triangulation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 228-240.
- HAYMAN, C. G., & TULVING, E. (1989b). Is priming in fragment completion based on a "traceless" memory system? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 941-956.
- HAYS, W. L. (1988). *Statistics*. New York: Holt, Reinhart & Winston.
- HINTZMAN, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, **87**, 398-410.
- HINTZMAN, D. L. (1990). Human learning and memory: Connections and dissociations. *Annual Review of Psychology*, **41**, 109-139.
- HINTZMAN, D. L. (1991). Contingency analyses, hypotheses, and artifacts: Reply to Flexser and Gardiner. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 341-345.
- HINTZMAN, D. L., & HARTRY, A. L. (1990). Item effects in recognition and fragment completion: Contingency relations vary for different subsets of words. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 955-969.
- JACOBY, L. L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning & Verbal Behavior*, **22**, 485-508.
- MARTIN, E. (1981). Simpson's paradox resolved: A reply to Hintzman. *Psychological Review*, **4**, 372-374.
- OSTERGAARD, A. L. (1992). A method for judging measures of stochastic dependence: Further comments on the current controversy. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 413-420.
- OSTERGAARD, A. L. (1994). Who is mistaken about priming in "recognition/identification" experiments? A reply to Tulving and Hayman. *European Journal of Cognitive Psychology*, **7**, 1-11.
- POSNER, M. I., & PETERSON, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, **13**, 25-42.

- PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T., & FLANNERY, B. P. (1992). *Numerical recipes in C* (2nd ed.). New York: Cambridge University Press.
- ROEDIGER, H. L., III, & McDERMOTT, K. B. (1993). Implicit memory in normal human subjects. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 63-131). New York: Elsevier.
- SCHACTER, D. L., COOPER, L. A., & DELANEY, S. M. (1990). Implicit memory for unfamiliar objects depends upon access to structural descriptions. *Journal of Experimental Psychology: General*, **119**, 5-24.
- SCHACTER, D. L., COOPER, L. A., DELANEY, S. M., PETERSON, M. A., & THARAN, M. (1991). Implicit memory for possible and impossible objects: Constraints on the construction of structural descriptions. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 3-19.
- SHIMAMURA, A. P. (1985). Problems with the finding of stochastic independence as evidence for multiple memory systems. *Bulletin of the Psychonomic Society*, **23**, 506-508.
- SLOMAN, S. A., HAYMAN, C. G., OHTA, N., LAW, J., & TULVING, E. (1988). Forgetting in primed fragment completion. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 223-239.
- TULVING, E. (1985). How many memory systems are there? *American Psychologist*, **4**, 385-398.
- TULVING, E., & HAYMAN, C. G. (1993). Stochastic independence in the recognition/identification paradigm. *European Journal of Cognitive Psychology*, **5**, 353-373.
- TULVING, E., SCHACTER, D. L., & STARK, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 336-342.
- WICKENS, T. D. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Erlbaum.
- WICKENS, T. D. (1993). Analysis of contingency tables with between-subjects variability. *Psychological Bulletin*, **113**, 191-204.
- WITHERSPOON, D., & MOSCOVITCH, M. (1989). Stochastic independence between two implicit memory tasks. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 22-30.

NOTES

1. Throughout the paper, small letters (*a, b*) will refer to probabilities and capital letters (*A, B*) will refer to frequencies (after Wickens, 1989).

2. Tulving and Hayman (1993) recently called into question Ostergaard's (1992) estimates of maximum dependence, claiming that he had used an incorrect (test-primed) baseline in the calculations. However, Ostergaard (1994) has defended these calculations, showing that the baseline proposed by Tulving and Hayman leads to some illogical conclusions about priming. Regardless of this issue about calculation, Tulving and Hayman endorsed the procedure as a way of examining stochastic independence.

3. Sloman et al. (1988) extensively reported the standard deviations of fragment completion performance. The median of the 47 reported standard deviation values was 0.15, and this value was thus used as a target value for the simulations.

4. This equation was derived by formulating the odds ratio in terms of Cell *a* probability and marginal probabilities:

$$\alpha = \frac{a(1+a-P_1-P_2)}{(P_1-a) \times (P_2-a)},$$

and solving for *a*. It is real valued for $\alpha \geq 0$, except for $\alpha = 1$, where it is undefined.

(Manuscript received August 31, 1994;
revision accepted for publication April 22, 1996.)