PERGAMON

Social Science & Medicine 52 (2001) 1843–1857

SOCIAL SCIENCE & MEDICINE

www.elsevier.com/locate/socscimed

Effect of orthographic and phonological similarity on false recognition of drug names

Bruce L. Lambert[a,b,*], Ken-Yu Chang[a], Swu-Jane Lin[a]

[a] *Department of Pharmacy Administration, University of Illinois at Chicago, 833 South Wood Street, Chicago IL 60612 7231, USA*
[b] *Department of Pharmacy Practice, University of Illinois at Chicago, 833 South Wood Street, Chicago IL 60612 7231, USA*

Abstract

Health professionals and patients tend to confuse drugs with similar names, thereby threatening patient safety. One out of four medication errors voluntarily reported in the US involves this type of drug name confusion. Cognitive psychology offers insight into how and why these errors occur. The objective of this investigation was to examine the effect of orthographic (i.e., spelling) and phonological (i.e., sound) similarity on the probability of making recognition memory errors (i.e., false recognitions). Prospective, computer-based, recognition memory experiments on 30 pharmacists and 66 college students were conducted. Participants viewed a study list of drug names and then a test list. The test list was twice as long as the study list and contained distractor names at progressively increasing levels of similarity to the study words. The task was to identify which test names were on study list and which were new. The main outcome measure was probability of making a false recognition error (i.e., of saying a new name was on the study list). Among pharmacists and college students, there was a strong and significant effect of similarity on the probability of making a false recognition error. It was concluded that both orthographic (i.e., spelling) and phonological (i.e., sound) similarity increase the probability that experts and novices will make false recognition errors when trying to remember drug names. Similarity is easily and cheaply measured, and therefore, steps should be taken to monitor and reduce similarity as a means of reducing the likelihood of drug name confusions.

Keywords: Drug names; Medication error; Patient safety; Pharmacists; Recognition memory; USA

Introduction

Drug therapy is the most common form of treatment offered by office-based physicians in the US, where nearly three billion prescriptions were dispensed in community pharmacies in 1999 (Woodwell, 1999; Anonymous, 1999). In order to reap the benefits and avoid the hazards of drug therapy, however, the right patient must receive the right drug in the right dose at the right time via the right route of administration. Too often, this does not occur (Cohen, 1999; Berwick & Leape, 1999; Lawrence, 1999; Allan & Barker, 1990; Kohn, Corrigan, & Donaldson, 2000). As a case in point; Patients sometimes receive the wrong drug because similarity in the spelling and/or pronunciation of drug names leads to errors in prescribing; Dispensing, and administration (e.g., *Xenical*® and *Xeloda*®, *cisplatin* and *carboplatin*, *amiodarone* and *amrinone*, *E-Vista*® and *Evista*®, *Celebrex*®, *Celexa*®, and *Cerebyx*®; *Dynacin*® and *DynaCirc*®, *Retrovir*® and *Ritonavir*®) (Davis, 1997). These errors are typically made by health professionals, but patients are not immune, especially when drug companies engage in so-called brand extensions, marketing multiple different sets of active

*Corresponding author. Department of Pharmacy Administration, University of Illinois at Chicago, 833 South Wood Street, Chicago IL 606 12 7231, USA. Tel.: +1-312-996-2411; fax: +1-312-996-6868.

E-mail address: lambertb@uic.edu (B.L. Lambert).

0277-9536/01/$ - see front matter
PII: S0277-9536(00)00301-4

ingredients under very similar brand names (see, e.g., *Tylenol Cold*®, *Tylenol Flu*®, as well as multiple *Chlor-Trimeton*® products, *Unisom*® products, *Mylanta*® products, *Maalox*® products, *Pepto-Bismol*® products, *Triaminic*® products, etc.) (US Pharmacopeia, 1996).

One out of four errors voluntarily reported in the US identifies name confusion as the primary cause (US Pharmacopeia, 1993, 1995, 1997; Davis, 1997). The lack of a mandatory reporting system makes it impossible to know the true rate of drug name confusions, but the enormous number of drugs being dispensed means that even an infinitesimally small error rate would cause a very high absolute number of errors. For example, if only 0.01% of the 3 billion prescriptions dispensed in 1999 involved a name confusion, 300,000 errors would still result! Orthographic (i.e., spelling) and phonological (i.e., sound) similarity between drug names are, in the jargon of quality control, *assignable* or *special* sources of variation; sources that can be measured objectively, monitored, and minimized in an effort to increase the quality and safety of medical care (Berwick, 1991).

In a series of case-control studies, lambert has shown that simple, automated measures of orthographic and phonological similarity are strongly associated with the commission of drug name confusion errors (Lambert, 1997; Lambert, Lin, & Gandhi, 1997; Lambert, Lin, Chang, & Gandhi, 1999). A measure of similarity based on the number of three-letter subsequences two names had in common was able to distinguish between cases (i.e., known errors) and closely matched controls with greater than 94% accuracy (Lambert et al., 1999). These results confirmed an association between similarity and name confusions; but they were based on retrospective analysis of an error database that was vulnerable to selection biases (Lambert et al., 1999). Moreover, the case-control studies were not based on an explicit model of the cognitive processes that caused the errors. Prior studies also did not differentiate between types of name confusion errors. Some may have been memory errors (i.e., errors in recall or recognition); some may have been perceptual errors (i.e., mistaken visual or auditory perception), and some may have been motor control errors (i.e., erroneous selection of an adjacent product from a pull-down computer menu). We felt we could draw stronger conclusions about causality from prospective experimental studies of a narrower class of errors. In the investigation reported below, We carried out experiments to examine one type of name confusion error, those involving recognition memory for visually presented, typewritten drug names. we used both pharmacists and lay people (college students) as participants, because both are known to be vulnerable to this type of confusion. The experiments were based on an explicit model of the cognitive processes that underlie recognition memory. We sought to answer two research questions.

RQ1: How does similarity affect pharmacists' recognition memory for visually presented drug names?
RQ2: How does similarity affect college students' recognition memory for visually presented drug names?

Theoretical background

Drug name confusions can result, at a minimum, from errors in memory, perception, and/or motor control (Reason, 1990). Here we focused only on recognition memory errors. (A parallel study of recall errors will be published separately, and a series of experiments on visual perception is underway.) The real-world task we were modeling is that of a health professional or patient who sees a medication name on an advertisement, package, printout, label, or computer screen, commits the name to memory, and then goes to retrieve the named medication from another location (e.g., a crash cart, shelf stock, drug store, etc.). When the professional or patient scans the shelf for the named medication, s/he is engaged in a recognition memory task, attempting to select the name from the shelf that matches the name stored in memory (Slack, 1991). When a false recognition occurs (i.e., when a similar name is selected rather than the target name), a medication error results. Our goal is to increase the safety and reliability of the drug use process by identifying and eventually minimizing the factors that cause false recognition errors.

Recognition memory

The standard recognition memory task has two parts — study and test. During study, the subject is presented with a list of words to be remembered. During test, the subject is presented with words from the study list mixed with new words (often called lures, foils or distractor words). The subject's task is to identify which words were presented on the study list (i.e. 'old' words) and which were not (i.e., 'new' words). An established set of phenomena has emerged from the hundreds of recognition memory experiments that have been done over the years. Among these are the list length effect (i.e., longer study lists produce more errors), the list strength effect (i.e., more study produces fewer errors), the word frequency effect (i.e., rare words are easier to recognize than frequent words), and the similarity effect (i.e., foils that are orthographically, phonologically, or semantically associated with study words are recognized falsely more frequently than unrelated foils) (Anderson, Bothell, Lebiere, & Matessa, 1998; Anderson & Lebiere,

1998; Underwood, 1970). The similarity effect was most relevant to the studies reported below.

The similarity effect was first observed when foils that were semantically associated to study words showed higher false recognition rates than unrelated controls (e.g., *bottom* was studied and *top* was falsely recognized) (Underwood, 1965), and it has since been repeatedly reproduced (Anisfeld & Knapp, 1968; Roediger & McDermott, 1995). The similarity effect has also been produced using foil words whose spelling and/or pronunciation were similar to study words (e.g., *sour* was studied and *slur* was falsely recognized, *maid* was studied and *made* was falsely recognized) (Nelson & Davis, 1972; Raser, 1972). This effect, which has clear implications for drug name confusion errors, has also been replicated many times (Wallace, 1968; Wallace, Stewart, & Malone, 1995a; Wallace, Stewart, Sherman, Heather, & Mallor, 1995b; Wallace, Stewart, Shaffer, & Wilson, 1998; Underwood & Zimmerman, 1973; Bencomo & Daniel, 1975; Sommers & Lewis, 1999).

Theories and models of recognition memory

In this section, we briefly summarize modern theories of recognition memory, focusing on explanations for false recognition errors. It is important to note, however, that memory experiments using word lists are the most common experiments in psychology, and the relevant literature is vast, spanning more than 100 years (Anderson et al., 1998). Interested readers should consult standard texts for a review (Anderson, 1995). In the last 30 years, several models have been proposed to account for well-known patterns of learning, remembering, and forgetting. Although these models differ in their details, they share many assumptions and make many similar predictions (Raaijmakers & Shiffrin, 1992).

The ACT-R model. Our analysis draws heavily on Anderson's Atomic Components of Thought-Rational (ACT-R) model, a recently proposed integration of existing models that has been successful in accounting for a wide range of list memory phenomena (Anderson et al., 1998; Anderson & Matessa, 1997; Anderson, Reder, & Lebiere, 1996; Anderson & Lebiere, 1998; Anderson, 1999; Anderson & Bower, 1974). ACT-R is an explicit, mathematical model, implemented as a set of computer programs. Only a brief verbal description of the theory is presented here. ACT-R states that cognition is supported by a set of procedural rules operating on a declarative memory. Declarative memory consists of an associative network of nodes, and each node encodes a fact. Each node has a level of activation based on its prior history of activation and on its current receipt of activation from associated nodes. Nodes have weighted links to one another. The strength of these links is a function of the likelihood that two nodes have been simultaneously activated in the past. The probability that a given chunk of knowledge will be retrieved from memory is a function of its activation, with the most active nodes having the highest probability of being retrieved (Anderson & Lebiere, 1998).

The ACT-R model of recognition memory is based on a simple rule:

IF the goal is to recognize if a word occurred in the list and there is a trace of the word in the list THEN accept it. (Anderson et al., 1998)

During the recognition task, this rule operates on a declarative memory representation of the list and the list context. (The list context refers to an episodic representation of one specific list, as opposed to any other memory representation of a given word.) The test word is used to probe memory in search of a memory trace that matches it. Words are typically represented as vectors of orthographic, phonological, and semantic features (Shiffrin & Steyvers, 1997; Hintzman, 1988). The level of activation of a word's memory trace is a function of its feature-by-feature similarity to the probe word (Hintzman, 1988).When subjects see a test word, they probe their memory for a trace of that word in the specific list context. If the probe reveals an active trace of the word, the word is identified as old. If not, it is identified as new. False recognitions occur when a new word partially matches an old word (i.e., when many but not all of the features in the new word match the trace of an old word). Since there is random noise in the activation values of each memory trace, words that partially match a trace can sometimes exceed the threshold for a correct match (Anderson et al., 1998). This is the basic framework for our investigations. Similarity increases the probability of false recognition errors because it increases the probability of partial matching between new and old words in the recognition memory task (Anderson et al., 1998; Anderson and Matessa, 1997; Anderson et al., 1996).

In this investigation we assumed that words were represented in memory as sets of orthographic and phonological features. We defined similarity in terms of the degree of overlap in these feature-sets, and then we examined the relationship between similarity and the probability of making a false recognition error. The project incorporated experiments to test the following hypotheses:

H1: False recognition errors will increase as orthographic similarity between target and foil names increases.

H2: False recognition errors will increase as phonological similarity between target and foil names increases.

Experiment 1: effect of orthographic similarity on pharmacists' false recognition of visually presented drug names

Methods

Design

This experiment was designed to examine the effect of orthographic similarity on pharmacists' short-term recognition of drug names, while holding a variety of other potentially confounding variables constant. Each participant viewed a 5-word study list and a 10-word test list. The task was to identify which words on the test list had appeared previously on the study list. The main variable of interest, orthographic similarity between study words and foils, was systematically varied. We expected recognition errors to be more likely as the similarity between study words and their corresponding foils increased. All of the experiments described below were approved in advance by the local institutional review board, and all participants orally consented to participate.

Participants

Fifteen licensed, practicing pharmacists participated in experiment 1. Participants were recruited from the clinical faculty and pharmacy resident staff of an academic medical center in the Midwest United States. All participants held the clinically oriented, Pharm.D. degree. Individuals were not paid for their participation.

Stimulus materials

Forty pairs of drug names were selected: 8 each at 5 progressively increasing levels of similarity (see Table 1). Orthographic similarity was measured by the bigram method with one space added to the beginning and ending of each word (Lambert, 1997; Stephen, 1994; Lambert et al., 1999). The first step in computing the bigram similarity between two words was to break the words into their two-letter subsequences. For example, the bigrams for the drug *Atarax*® were {*_a, at, ta, ar, ra, ax, x_*}. The bigrams for the drug *Marax*® were {*_m, ma, ar, ra, ax, x_*}. The Dice coefficient was then used to compute a similarity score between 0 and 1

$$\text{OrthoSim} = 2C/(B + A),$$

where A was the number of bigrams in the first word, B was the number of bigrams in the second word, and C was the number of bigrams that occur in both words. Atarax and Marax shared 4 bigrams {*ar, ra, ax, x_*}. Their bigram similarity score was $(2*4)/(7+6) = 0.615$.

Eight sets of 5 pairs, one pair at each level of similarity, were blocked by frequency to prevent frequency from confounding the effects of similarity in the memory (see Table 1). Names and prescribing frequencies were taken from the combined 1992–1994 National Ambulatory Medical Care Survey (NAMCS) databases (US Department of Health and Human Services, 1996; Schappert, 1994; Woodwell & Schappert, 1995; Schappert, 1996). Both brand and generic names appeared in the NAMCS data, but our stimuli ended up including only 6 generic names. The eight sets of name pairs permitted construction of 16 trials, with each word appearing in one trial as a target word and in one trial as a foil.

Table 1
Stimulus materials for recognition memory task (orthographic similarity) (experiments 1 and 3)[a]

Log frequency	Bigram similarity	Names	
5.94	0.78	Prolixin®	Procolin®
5.94	0.47	Tenuate®	Trilisate®
5.93	0.21	Zithromax®	Capitrol®
5.93	0.13	Rogenic®	Benylin®
5.93	0.00	Aclovate®	Nicobid®
3.94	0.75	Aramine®	Anamine®
3.92	0.47	Trantoin®	Triacin®
3.91	0.21	Nephramine®	Sinophen®
3.92	0.13	Dialose®	Pinoval®
3.92	0.00	Paraflex®	Otrivin®
4.86	0.74	Hydrocort®	Hydrocet®
4.86	0.47	Antiminth®	Timentin®
4.86	0.21	Minizide®	Allergine®
4.86	0.13	Hexalol®	Temaril®
4.86	0.00	Belexal®	Marazide®
5.75	0.71	Urisep®	Urised®
5.75	0.44	Exelderm®	Eldepryl®
5.75	0.21	Dermatop®	captopril
5.75	0.13	Mylicon®	Empirin®
5.75	0.00	Senokot®	Efudex®
5.20	0.71	rifampin	Rifadin®
5.20	0.44	Choledyl®	Cholybar®
5.20	0.21	Trinsicon®	Atabrine®
5.20	0.13	Genora®	Desferal®
5.20	0.00	Claforan®	Merital®
5.34	0.67	Pramasone®	Orasone®
5.35	0.47	Drixoral®	Fluoral®
5.35	0.21	Enduron®	dantrolene
5.35	0.13	Glucola®	Talacen®
5.35	0.00	Rowasa®	Ferralet®
6.58	0.67	Isordil®	Isomil®
6.59	0.40	Indocin®	doxepin
6.56	0.21	Antivert®	Ascriptin®
6.58	0.13	Zoladex®	Relafen®
6.58	0.00	Lotensin®	Nizoral®
5.12	0.67	Panadol®	nadolol
5.13	0.44	halothane	Loxitane®
5.11	0.21	Theraplex®	Hexadrol®
5.12	0.13	Estinyl®	Vepesid®
5.12	0.00	Betalin®	Rynatuss®

[a] *Note*: Log frequency refers to the average frequency of the two words in a pair.

Procedures

Experiments were conducted using the SuperLab® experiment program for the Macintosh computer (Cedrus, 1991). The recognition task consisted of one practice trial and 16 experimental trials. Each trial consisted of a study phase and a test phase. During the study phase, 5 names were displayed on the computer monitor at a rate of one name per second. After the fifth name was presented, a message appeared alerting the participants that the test phase was about to begin. During the test phase, 10 names were presented. The 10 names in the test set were comprised of 5 names from the study list and 5 new names which had not appeared on the study list. Words on the test list were presented in a random sequence. The task was to indicate, for each name on the test list, whether it had appeared on the study list or not. Participants indicated their decision by pressing one of two different keys on the computer keyboard. Participants took as much time as they needed to make the recognition judgment. Each participant produced 160 responses (i.e., 16 trials, 10 test words per trial), only 80 of which could have been false positive responses (i.e., only 80 responses were responses to foils). After the experiment was completed, participants were asked to rate the subjective familiarity of each name in the experiment on a seven-point, semantic differential scale that ranged from -3 (i.e., extremely unfamiliar) to $+3$ (i.e., extremely familiar).

Analysis plan

The dichotomous dependent variable in this experiment was false positive recognitions. Foil words correctly recognized as new (i.e., not from the study list) were scored as 0, and foil words falsely recognized as old (i.e., from the study list) were scored as 1. The independent variable was: (a) orthographic similarity, a continuous variable reflecting the bigram similarity for each target–foil pair. The control variables (i.e., those that were included to control nuisance variation and to rule out alternative hypotheses) were (a) target frequency, a continuous variable reflecting the log (base 10) of the NAMCS prescribing frequency of the target word; (b) foil frequency, a continuous variable reflecting the log of the NAMCS prescribing frequency of the foil word; (c) target familiarity, an ordinal variable representing the subjective familiarity of the target name; (d) foil familiarity, an ordinal variable representing the subjective familiarity of the foil name; (e) lag, an ordinal variable representing the number of names intervening between presentation of target and foil; (f) trial, an ordinal variable representing the sequential position of a given response within the set of 160 responses; and (g) target, a dichotomous variable scored as 1 if the current foil name had appeared on a previous list as a target and 0 otherwise.

Data were analyzed using MIXOR, a system for doing mixed effects, logistic regression modeling of dichotomous and ordinal data. The mixed-effects logistic regression model accommodates nesting of experimental conditions within subjects for a binary outcome and a mixture of discrete and continuous covariates that can vary either at the level of the subject or the experimental condition (Hedeker & Gibbons, 1996, 1994; Hedeker, 1999).

Our modeling strategy included multiple steps. The first step was to identify the correct scale for each independent and control variable. We did this by separately plotting the log odds of error as a function of each independent or control variable. If these plots were linear, terms were entered as simple linear terms. If the plot revealed an obvious nonlinearity, we selected a scale to fit the nonlinear form of the function (Hosmer & Lemeshow, 1989; Selvin, 1996). In this case, we primarily considered quadratic terms. Having identified the appropriate scale for each independent and control variable, we used Kleinbaum's method of backward elimination to decide which variables to include in the final model (Kleinbaum, 1994). According to this method, the analyst begins with a full model and then proceeds to eliminate as many terms as possible, using likelihood ratio tests (analogous to partial F-tests in ordinary least-squares regression) to decide which terms contribute significantly to the model's fit. Higher order terms (e.g., interaction terms, squared terms) are eliminated first, then first-order terms. Interested readers should consult the text for a complete description of the approach (Kleinbaum, 1994).

The final step in our modeling strategy was to assess goodness-of-fit. Unlike the case of ordinary least-squares regression, where R^2 provides a widely agreed-upon measure of fit, in logistic regression, there is no consensus measure of goodness-of-fit (Pedhazur, 1997; Hosmer & Lemeshow, 1989; Kleinbaum, 1994; Selvin, 1996). Rather, multiple measures are available. For each model, we report several different indices of goodness-of-fit: classification accuracy, Hosmer–Lemeshow's C-test based on deciles of risk, and fit between observed and predicted probabilities of error at selected levels of similarity.

To calculate classification accuracy, we imposed a threshold on predicted probability scores to generate classifications (e.g., if predicted probability >0.5, then classify as error). Because we viewed sensitivity and specificity as equally important, we searched for the threshold that minimized the difference between sensitivity (the true positive rate) and specificity (the true negative rate). We reported sensitivity, specificity and overall accuracy at the selected threshold (Hosmer & Lemeshow, 1989).

For the Hosmer-Lemeshow test, we sorted observations into deciles of risk using their predicted probability

of error as the sort key. We then compared the observed and expected number of errors and correct responses within each decile of risk (Hosmer & Lemeshow, 1989; Selvin, 1996). Hosmer and Lemeshow's C is a chi-square statistic with 8 degrees of freedom (when deciles are used); the null hypothesis is that the data come from the same distribution (i.e. that the model fits). Plots of predicted versus observed frequencies are provided for each experiment. All statistical tests used an alpha = 0.05 unless otherwise noted.

The third and final measure of fit graphically compared observed and predicted probabilities of error at selected levels of similarity (5 levels in experiments 1 and 3, 4 levels in experiments 2 and 4). To compute these probabilities, we evaluated the fitted model with all control variables set to their mean values and similarity set to the average value within a given level.

Results

Descriptive statistics for the variables in experiment 1 are given in Table 2. Fig. 1 provides a graphical display of the results. Overall, 87.16% of the foils were correctly recognized as 'new' (i.e., not from the study list), while 12.83% were incorrectly recognized as 'old' (i.e., from the study list). Following the backward elimination strategy, three logistic regression models were fitted to the data. The first included the intercept, control variables and independent variables, including squared terms for trial and similarity. The quadratic term for trial was eliminated from this model, based on an insignificant change in the likelihood ratio test comparing models with and without it (Kleinbaum, 1994). The quadratic similarity term was then eliminated in the same manner. The final model included all of the control variables plus similarity. We did not try to eliminate first-order control variables, preferring instead to leave them in to control for confounding (Kleinbaum, 1994). Parameter estimates for the final model are given in Table 3. As predicted, false recognition errors increased as similarity increased ($b = 2.73$, $p < 0.0001$). In addition, the effect of trial was significant ($b = -0.01$, $p < 0.01$), with participants performing slightly better in later trials than in earlier trials. No other variables were significant in this model.

Table 2
Descriptive statistics for experiment 1[a]

Variable	Minimum	Maximum	Mean	SD
False pos. error	0.00	1.00	0.13	0.33
Similarity	0.00	0.78	0.30	0.25
Trial	2.00	160.00	80.30	46.40
Log frequency	2.46	6.88	5.01	1.01
Lag	1.00	14.00	7.30	3.43
Familiarity	−3.00	3.00	0.25	2.46

[a] Note: Log frequency and familiarity data are listed only once, even though, conceptually, the design refers to target frequency and foil frequency (or target familiarity and foil familiarity). Frequency and familiarity ratings of words were the same whether they were seen in a given trial as target or foil. This table also gives descriptive statistics for experiment 3, except for familiarity scores, which are given in the text for experiment 3.

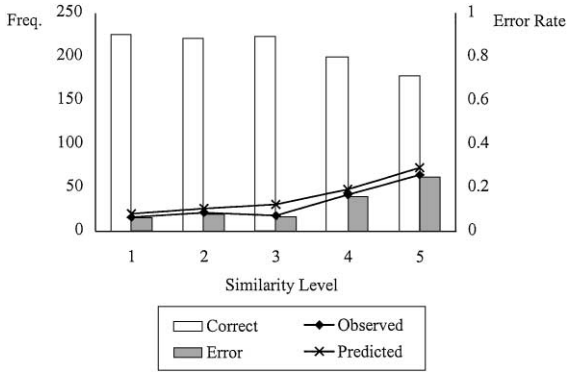

Fig. 1. Pharmacists' recognition performance for orthographically similar names (experiment 1). Freq. refers to the observed frequency of correct or incorrect responses. Light bars represent correct responses; dark bars represent errors. The trend lines illustrate the observed and predicted increases in the false positive error rate. Although similarity was a continuous variable in our analyses, we have divided it into 5 ranges for the purposes of creating these illustrations. Levels 1–5 corresponded to mean bigram similarity values of 0.00, 0.13, 0.21, 0.45, and 0.71, respectively.

The model had sensitivity of 66.8%, sensitivity of 67.1%, and overall accuracy of 67.1% when a predicted probability threshold of 0.118 was used for classification. The Hosmer-Lemeshow test showed no evidence of lack of fit ($X^2(8) = 6.12$, $p = 0.4$). Panel (a) of Fig. 2 shows a plot of predicted versus observed error frequencies at each decile of risk. The points follow the 45 degree diagonal from the origin, indicating a good fit. Fig. 1 shows the fit between observed and predicted probabilities at 5 levels of similarity.

Discussion of experiment 1

The results of experiment 1 supported our first hypothesis. The log odds of making a false recognition error increased significantly as orthographic similarity increased. The more orthographic features shared by two words, the higher the likelihood that the foil word would partially match to the target during recognition. There was also a significant effect of trial on performance, with participants making fewer errors on later than on earlier trials. We attributed this improvement to the effects of practice, especially because participants

were only given one practice trial to familiarize themselves with the task. Although it tended to predict more errors than were observed (see Fig. 1), a simple, 8-variable model provided a good fit with the observed data, suggesting that our measure of orthographic similarity was valid and that no important variables were missing from the model.

Table 3
Experiment 1: parameter estimates for logistic regression model predicting false positive recognition errors[a]

Variable	Estimate	SE	Z
Intercept	−1.70	2.49	−0.68
Similarity	2.73	0.61	4.48[b]
Trial	−0.01	0.004	−2.50[c]
Log stim. freq.	−0.14	0.24	−0.58
Log foil freq.	−0.10	0.29	−0.34
Lag	0.06	0.04	1.50
Target familiarity	−0.08	0.07	−1.14
Foil familiarity	−0.06	0.12	−0.50
Target	0.61	0.35	1.74

[a] Freq. = frequency, −2 log likelihood = 809.06.
[b] $p < 0.0001$.
[c] $p < 0.01$.

Experiment 2: effect of phonological similarity on pharmacists' false recognition of visually presented drug names

Methods

Design

Experiment 2 was designed to parallel experiment 1, the main difference being that similarity in this experiment was measured in terms of phonological rather than orthographic features.

Participants

Fifteen licensed, practicing pharmacists participated in experiment 2. This was a different group than the 15 who participated in experiment 1, although they were recruited from the same academic medical center. All participants held the clinically oriented, Pharm.D. degree. Individuals were not paid for their participation.

Stimulus materials

Stimulus materials consisted of 32 pairs of generic drug names, 8 each at 4 progressively increasing levels of similarity (see Table 4). To be included in this

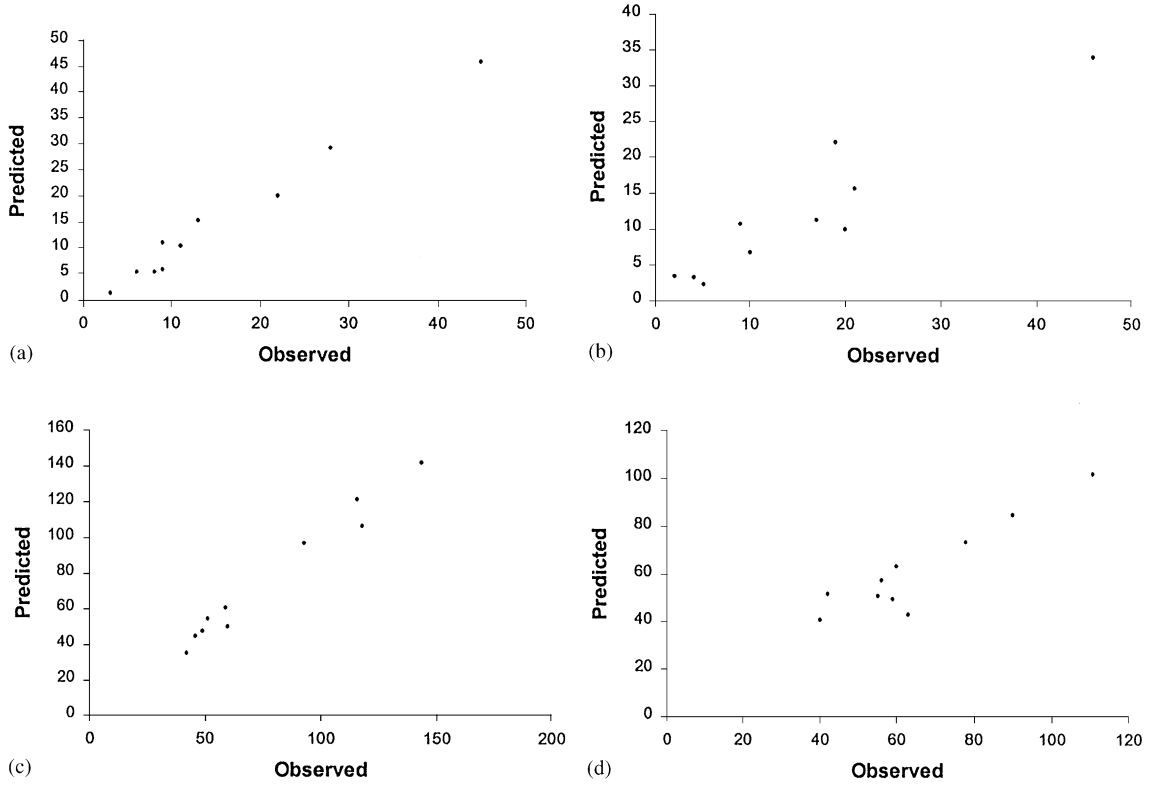

Fig. 2. Predicted versus observed frequency of false recognition errors at 10 deciles of risk: (a) experiment 1, (b) experiment 2, (c) experiment 3, (d) experiment 4. A perfect fit would be indicated by a scatter of points along the 45° diagonal from the origin. (see Hosmer & Lemeshow, 1989for details on computing fit using deciles of risk.).

experiment, names had to be listed both in the NAMCS database and in the *USP Dictionary of USAN and International Drug Names* (US Pharmacopeia, 1998) because prescribing frequency data were taken from the NAMCS database, and pronunciation guides were taken from the *USP Dictionary*.

An ad hoc measure of phonological similarity was developed specifically for this experiment. Several phonological characteristics had been identified as important in previous research on similarity and memory (Drewnowski & Murdock, 1980). These included number of syllables, location of stressed syllable, initial syllable, terminal syllable, and stressed vowel. Based on these features, similarity was defined as follows:

$$\text{Phono Sim}(\text{word}_i, \text{word}_j) = 0.5\left(\frac{2C}{B+A}\right) + 0.5\frac{(2D+E+F+G+H)}{6},$$

where A was the number of syllables in one word, B was the number of syllables in the other word, C was the number of common syllables, D was a binary feature representing a match between initial syllables, E was a binary feature representing a match between terminal syllables, F was a binary feature representing a match between accented syllables, G was a binary feature representing a match between accent positions, and H was a binary feature representing a match between number of syllables. This ad hoc measure gave half of the weight to the commonality in syllables and half of the weight to specific phonological characteristics, with initial phoneme getting twice the weight as the other phonological features (Wallace et al., 1995a,b, 1998).

To compute the phonological similarity between lincomycin and tobramycin, we began by retrieving the respective pronunciation guides from the USP Dictionary: (*lin koe mye*' *sin*) and (*toe bra mye*' *sin*). An apostrophe indicated the accented syllable. These names shared two syllables {*mye*', *sin*}. They had different initial syllables ($D=0$). They had the same accented syllable ($E=1$) in the same accent position ($F=1$). They had the same number of syllables ($G=1$) and the same terminal syllable ($H=1$). Thus, the similarity between these two names was calculated as follows:

$$\text{Phono Sim}(\text{word}_i, \text{word}_j) = 0.5\left(\frac{2\times2}{4+4}\right) + 0.5\frac{(2\times0+1+1+1+1)}{6} = 0.58.$$

Sets of names at varying levels of similarity were again blocked by NAMCS prescribing frequency so that there would be no correlation between similarity and frequency (US Department of Health and Human Services, 1996).

Table 4
Stimulus materials for phonological recognition task (phonological similarity) (experiments 2 and 4)[a]

Log frequency	Similarity	Names	
4.55	0.75	Chloroform	Chloroquine
4.76	0.42	Glycerin	Tolmetin
4.41	0.25	Cisapride	Urea
4.67	0.00	Benzoin	Filgrastim
5.84	0.71	Acyclovir	Ganciclovir
5.84	0.38	Felodipine	Nifedipine
5.91	0.17	Isosorbide	Oxytocin
5.81	0.00	Aminophylline	Baclofen
5.86	0.67	Betamethasone	Dexamethasone
5.55	0.43	Mephobarbital	Metronidazole
5.52	0.19	Griseofulvin	Riboflavin
5.72	0.00	Cyclosporine	Nitrofurantoin
4.98	0.67	Tolazamide	Tolbutamide
5.02	0.42	Mannitol	Sorbitol
5.00	0.17	Melphalan	Propofol
4.97	0.00	Phenol	Probenecid
5.83	0.58	Atropine	Loxapine
5.78	0.42	Calamine	Phentermine
5.80	0.17	Captopril	Ipecac
5.79	0.00	Dapsone	Meprobamate
4.97	0.58	Digitalis	Digitoxin
4.91	0.38	Didanosine	Dienestrol
4.93	0.17	Glucagon	Ichthammol
4.91	0.00	Flutamide	Pentoxifylline
6.25	0.50	Amikacin	Bacitracin
6.40	0.38	Ceftazidime	Cephalexin
6.31	0.21	Methotrexate	Ofloxacin
6.18	0.00	Indomethacin	Nystatin
5.19	0.58	Cefaclor	Cephradine
5.27	0.42	Carbachol	Carmustine
5.26	0.17	Lactulose	Succimer
5.23	0.00	Estrone	Misoprostol

[a] *Note:* Log frequency refers to the average frequency of the two words in a pair.

Procedures and analysis plan

The procedures were the same as those used in Experiment 1, except in this experiment, each study list was comprised of 8 words and each test list was comprised of 16 words. Each participant produced 128 responses (i.e., eight trials, 16 test words per trial), only 64 of which could have been false positive responses (i.e., only 64 responses were responses to foils). The analysis plan was essentially the same as in experiment 1.

Results

Descriptive statistics for the variables in Experiment 2 are given in Table 5. Fig. 3 displays the results graphically. Overall, 84.06% of the foils were correctly recognized as new (i.e., not from the study list), while 15.93% were falsely recognized as old (i.e., from the

Table 5
Descriptive statistics for experiment 2[a]

Variable	Minimum	Maximum	Mean	SD
False pos. error	0.00	1.00	0.16	0.37
Similarity	0.00	0.75	0.30	0.24
Trial	4.00	128.00	65.29	35.94
Log Freq.	2.45	6.70	5.07	0.92
Lag	1.00	23.00	12.80	5.31
Familiarity	−3.00	3.00	2.05	1.55

[a] *Note*: Freq. = frequency. Log frequency and familiarity data are listed only once, even though, conceptually, the design refers to target frequency and foil frequency (or target familiarity and foil familiarity). Frequency and familiarity ratings of words were the same whether they were seen in a given trial as target or foil. This table also gives descriptive statistics for experiment 4, except for familiarity scores, which are given in the text for experiment 4.

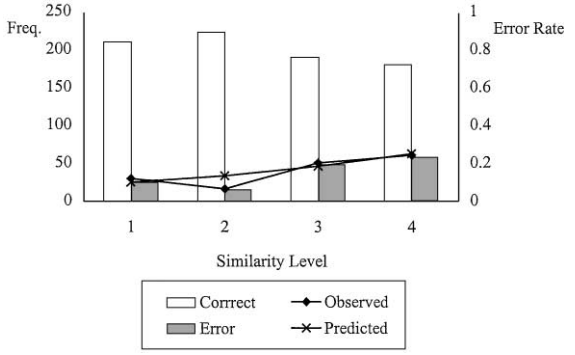

Fig. 3. Pharmacists' recognition performance for phonologically similar names (experiment 2). Freq. refers to the observed frequency of correct or incorrect responses. Light bars represent correct responses; dark bars represent errors. The trend lines illustrate the observed and predicted increases in the false positive error rate. Although similarity was a continuous variable in our analyses, we have divided it into 4 ranges for the purposes of creating these illustrations. Levels 1–4 corresponded to mean bigram similarity values of 0.00, 0.19, 0.40, and 0.63, respectively.

study list). Based on plots of log odds of error against each independent and control variable, quadratic terms for trial, target familiarity, foil familiarity, and similarity were evaluated in addition to the standard set of controls and similarity. All quadratic terms except similarity-squared were eliminated as the result of insignificant likelihood ratio tests. Comparison of models including all controls plus similarity and similarity-squared, controls plus similarity, and controls plus similarity-squared, were not statistically distinguishable, though each was significantly better than a model that included only control variables. Thus, the simplest model, containing all first-order control variables plus similarity, was chosen as the final model. Table 6 displays parameter estimates for the final model. Similarity had a significant effect on false recognitions ($b = 2.17$, $p < 0.05$). False recognition errors were more likely to occur as similarity increased. Foil names presented on previous trials as targets were more likely to be falsely recognized than foil names not previously presented ($b = 1.86$, $p < 0.0001$). The effect of trial was significant ($b = -0.01$, $p < 0.01$), with participants performing slightly better in later trials than in earlier trials. No other variables were significant in this model.

The model had 62.1% sensitivity, 65.3% specificity, and 64.8% overall accuracy when a threshold of 0.114 was used for classification. The Hosmer-Lemeshow *C*-test indicated that the fit was not optimal, ($X^2(8) = 29.56$, $p < 0.01$). At mid and high deciles of risk, the model predicted too few errors. Panel (b) of Fig. 2 shows a scatterplot of predicted versus observed error frequencies at the 10 deciles of risk.

Discussion of experiment 2

Results of experiment 2 supported our second hypothesis. Pharmacists' false recognition errors increased as phonological similarity increased. These results suggested that our measure of phonological similarity, although somewhat ad hoc, was valid. The significance of the target variable indicated that participants often fell victim to source monitoring errors. That is, when a foil word had been presented on a previous trial as a target, participants incorrectly identified it as having been "old". This was not an effect of similarity, but rather of confusion between the current study list and previously presented study lists. (We are grateful to an anonymous reviewer for pointing out the possibility of such source-monitoring errors.) It

Table 6
Experiment 2: Parameter estimates for logistic regression model predicting false positive recognition errors[a]

Variable	Estimate	SE	Z
Intercept	−4.04	2.23	−1.81
Similarity	2.17	0.97	2.23[b]
Trial	−0.01	0.01	−2.00[c]
Log target freq.	0.30	0.20	1.50
Log foil freq.	−0.22	0.21	−1.05
Lag	0.05	0.03	1.67
Target familiarity	0.03	0.14	0.21
Foil familiarity	0.01	0.19	0.05
Target	1.86	0.42	4.43[d]

[a] Freq. = frequency, −2 log likelihood = 730.15.
[b] $p < 0.05$.
[c] $p < 0.01$.
[d] $p < 0.0001$

is important to note, however, that the effect of similarity was still significant, even when the confounding effects of source monitoring errors were controlled. The significant effects of trial again showed that participants benefited from practice. Both the classification data and the Hosmer-Lemeshow test indicated a lack of fit between model and data. For example, we observed but could not explain an apparent dip in the error rate at the second level of similarity (see Fig. 2). The lack of fit suggests either that our measure of similarity was flawed or that the underlying pronunciation guides were inappropriate. We speculate that this lack of fit was due to the somewhat informal and ad hoc nature of our measure of phonological similarity. Subsequent research should strive to improve the measurement of phonological similarity.

Experiment 3: effect of orthographic similarity on college students' false recognition of visually presented drug names

Methods

Design and participants

Experiment 3 was designed to be identical to experiment 1 except with respect to the sample. The sample for this experiment was larger and included no health professionals. Participants in Experiment 3 were 33 college students. The majority of students were undergraduate psychology majors who participated in exchange for course credit. Students who received course credit for participation were not eligible for payment. A small number of participants were recruited from the general student population. These students were each paid $10 for their participation. Payment was primarily an incentive to get students to come to the medical campus, approximately two miles west of the main campus.

Stimulus materials, procedures, and analysis plan

Stimulus materials were the same as those used in experiment 1 (see Table 1). The experimental task and statistical analysis were also the same as in experiment 1.

Results

Since the same stimuli were used in both experiments 1 and 3, descriptive statistics for similarity, trial, log frequency, and lag for both experiments can be found in Table 2. The mean familiarity of the names for the participants in experiment 3 was −2.19 (SD = 1.61). In all, 70.53% of the foils were correctly recognized as new (i.e., not from the study list), while 29.47% were incorrectly recognized as old (i.e., from the study list). A graph of the results is given in Fig. 4. Table 7 displays parameter estimates for the logistic regression model. Increases in similarity ($b = 2.38$, $p < 0.0001$), trial ($b = -0.01$, $p < 0.01$), lag ($b = 0.03$, $p = 0.01$), target familiarity-squared ($b = -0.04$, $p < 0.05$), and trial-squared ($b = 0.06$, $p < 0.05$) were significantly associated with changes in the probability of false positive recognition errors. As predicted, when similarity between target and foil increased, so did the probability of false positive recognition errors. The effects of trial were complex and nonlinear. The larger the lag between

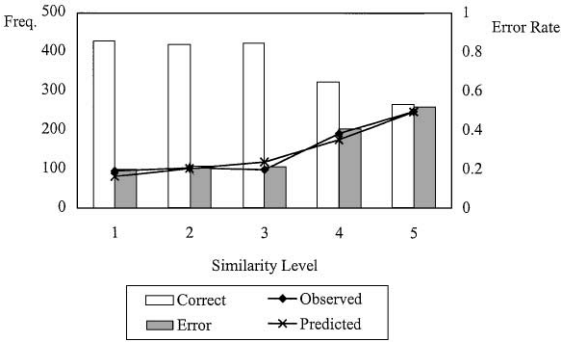

Fig. 4. College students' recognition performance for orthographically similar names (experiment 3). Freq. refers to the observed frequency of correct or incorrect responses. Light bars represent correct responses; dark bars represent errors. The trend lines illustrate the observed and predicted increases in the false positive error rate. Although similarity was a continuous variable in our analyses, we have divided it into 5 ranges for the purposes of creating these illustrations. Levels 1–5 corresponded to mean bigram similarity values of 0.00, 0.13, 0.21, 0.45, and 0.71, respectively.

Table 7
Experiment 3: parameter estimates for logistic regression model predicting false positive recognition errors[a]

Variable	Estimate	Stand. Error	Z
Intercept	−0.95	0.49	−1.93
Similarity	2.38	0.26	9.15[b]
Trial	−0.01	0.004	−2.50[c]
Log target freq.	−0.02	0.05	−0.40
Log foil freq.	−0.05	0.07	−0.71
Lag	0.03	0.01	3.00[d]
Target familiarity	0.01	0.06	0.17
Foil familiarity	0.04	0.05	0.80
Target	0.33	0.22	1.50
Target familiarity^2	−0.04	0.02	−2.00[d]
Trial^2	0.06	0.03	2.00[d]

[a] Freq. = frequency. Trial^2 was divided by 1000 before being entered in the model, −2 log likelihood = 2932.33.
[b] $p < 0.0001$.
[c] $p < 0.01$
[d] $p < 0.05$.

presentation of target and foil, the more likely was a false positive error. As the square of target familiarity increased, the probability of a false recognition decreased.

The model had 64.2% sensitivity, 63.2% specificity, and 63.5% overall accuracy when a threshold of 0.28 was used for classification. The Hosmer–Lemeshow *C*-test gave no evidence of lack of fit, ($X^2(8) = 7.24$, $p = 0.4$). Panel (c) of Fig. 2 shows a scatterplot of predicted versus observed error frequencies at the 10 deciles of risk, indicating a very good fit.

Discussion of experiment 3

The results of experiment 3 supported our first hypothesis and, in large part, mirrored the results observed in experiment 1. For both pharmacists and college students, increased orthographic similarity was associated with a significant increase in the log odds of a false recognition. The results suggested that excessive similarity between drug names posed a risk not only to health professionals as they handled drugs, but also to patients, who may report taking the wrong drug when interviewed by health professionals or select the wrong over-the-counter drug in a retail setting. That is, similarity between non-prescription drug names, which are selected and administered by consumers themselves, may cause errors in self-care. The significant effect of lag on false recognitions was in all likelihood caused by the slow decay of memory traces over time, an effect that has been well-documented in the literature (Anderson et al., 1998). As time passed, memory traces for studied words slowly decayed, thereby increasing the chances of a partial match to a similar foil word. Trial and target familiarity had complex, nonlinear effects on the probability of a false recognition error. We offer no interpretation of these effects, except to note that they are nonzero and must be controlled in order to examine the main effect of similarity. As in experiment 1, the fit between model and data was good (see Figs. 2 and 4).

Experiment 4: effect of phonological similarity on college students' false recognition of visually presented drug names

Methods

Design and participants

This experiment was designed to replicate experiment 2, using college students as participants rather than pharmacists. Thus, the design, stimulus materials, procedures, and analysis plan were the same as experiment 2 (see Table 4). Participants in experiment 4 were 33 college students (not the same group from experiment 3). Most were undergraduate psychology majors who participated in exchange for course credit. Several were recruited from the general student population. These students were each paid $10 for their participation.

Results

Since the same stimuli were used in both experiments 2 and 4, descriptive statistics for similarity, trial, log frequency, and lag for both experiments can be found in Table 5. The mean familiarity of the names for the participants in experiment 4 was –1.81 (SD = 1.95). Overall, 69.03% of the foils were correctly recognized as new, and 30.97% of the foils were incorrectly recognized as old. Fig. 5 displays the results graphically. Based on bivariate plots, squared terms for trial, target frequency, foil frequency, lag, and similarity were evaluated along with the standard set of controls and independent variables. All squared terms except similarity-squared were excluded from the model based on insignificant likelihood ratio tests. Table 8 shows parameter estimates for the final logistic regression model. The regression of error probability on similarity was quadratic, with errors increasing over most of the range of similarity and decreasing at very high levels of similarity (see Table 8 for details). Lag ($b = 0.04$, $p < 0.01$) was also significantly associated with increased probability of false positive errors. As the lag between target and foil increased, so did the likelihood of a false positive recognition error. Foil names that had appeared on previous trials as targets were more likely to be falsely recognized than names seen initially as foils ($b = 0.52$, $p < 0.01$).

The model had 58.9% sensitivity, 57.9% specificity, and 58.2% overall accuracy when a threshold of 0.29 was used for classification. The Hosmer–Lemeshow C test indicated that the fit was not optimal, ($X^2(8) = 18.69$, $p < 0.05$). At low deciles of risk, the model predicted too few errors. Panel (d) of Fig. 2 shows a scatterplot of predicted versus observed error frequencies at the 10 deciles of risk.

Discussion of experiment 4

The results of experiment 4 were partially supportive of hypothesis 2 and were generally consistent with what we observed in experiments 1–3. The log odds of a false recognition was a quadratic function of phonological similarity between target and foil, increasing at low and medium levels of similarity, but decreasing at very high levels. We have refrained from offering a speculative explanation of this quadratic effect. As in experiment 3, longer lags were associated with higher false recognition probabilities. We interpreted these results as reflecting the effects of decaying memory traces (Anderson et al., 1998). As in experiment 2, participants were vulnerable to source monitoring errors, such that foil names that had previously been presented as targets were more likely to be falsely recognized than names that had not been previously presented. Participants clearly had some

difficulty remembering whether a given name had been on the current study list or a previously presented study list. As in experiment 2, there was evidence of lack of fit between model and data, especially where the observed error rate declined at the second level of similarity (see Figs. 2 and 5), suggesting problems with the phonological similarity measure and/or the absence of important additional variables.

General discussion

The benefits of modern, technologically intensive, medical care have come at a high price in terms of preventable harm to patients. Based on recent estimates, medical error is among the leading causes of death in developed countries (Berwick & Leape, 1999; Lawrence, 1999; Bates et al., 1995; Kohn et al., 2000). Although it may be tempting to blame individual doctors, nurses, or pharmacists for errors, an emerging consensus suggests that such a 'culture of blame' is a significant obstacle to the reduction of medical error. Instead, experts in error reduction and quality control emphasize a systems approach to the analysis and prevention of medical errors (Leape et al., 1995; Reason, 1990). This approach asserts that errors are primarily the result of poorly designed systems (e.g., poor lighting, poorly organized tasks, noise, distractions, similar names). The key to prevention, from a systems viewpoint, is to design systems in such a way that errors are either difficult or impossible to make. Systems theorists also emphasize the need for ongoing monitoring and analysis of system inputs and outputs as a means of identifying underlying weaknesses in system design (Leape et al., 1995).

Leape and colleagues identified failures in dose and identity checking as the second most common system failure in their analysis of 334 errors. Most of these errors were attributed to similarities in packaging and drug nomenclature (Leape et al., 1995). We agree that excessive similarity between drug names is a frequent proximal cause of medication errors, and we assert that the existence of so many similar pairs of names reflects systemic flaws in the way drug names are evaluated and approved (Lambert et al., 1999). The drug name approval process is prone to error, in part because it involves multiple organizations that apply multiple conflicting criteria. The other notable weakness in current processes for name approval is the reliance on subjective judgments about similarity (Boring, 1997a, b; Boring, Homonnay-Weikel, Cohen, & Di Domizio, 1996; Boring, Stein, & Domizio, 1998). Here and elsewhere, we have shown that similarity is an assignable source of variation in the drug name confusion error rate (Lambert, 1997; Lambert et al., 1999; Berwick, 1991). Simple automated measures are available to make objective assessments of similarity between drug names (Stephen, 1994). The availability of such measures makes it difficult to justify ongoing reliance on subjective judgments as the basis for drug name approval decisions. These technologies should enable name approval agencies to ensure that new drug names are surrounded by a 'zone of safety' in the overall space of names. Preventing new names from violating the zone of safety of existing names should reduce the incidence of drug name confusion errors.

In this study, we have gone beyond retrospective analysis of case reports to offer an explicit model of the cognitive processes that lead to one type of memory

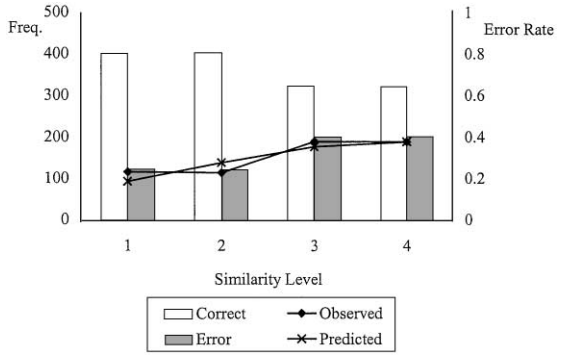

Fig. 5. College students' recognition performance for phonologically similar names (experiment 4). Freq. refers to the observed frequency of correct or incorrect responses. Light bars represent correct responses; dark bars represent errors. The trend lines illustrate the observed and predicted increases in the false positive error rate. Although similarity was a continuous variable in our analyses, we divided it into 4 ranges for the purposes of creating these illustrations. Levels 1–4 corresponded to mean bigram similarity values of 0.00, 0.19, 0.40, and 0.63, respectively.

Table 8
Experiment 4: parameter estimates for logistic regression model predicting false positive recognition errors[a]

Variable	Estimate	Stand. Error	Z
Intercept	−1.81	0.68	−2.66[b]
Similarity	1.73	0.23	7.46[c]
Similarity^2	−2.88	1.21	−2.36[d]
Trial	−0.00	0.00	−0.3
Log target freq.	0.01	0.06	0.13
Log foil freq.	0.04	0.07	0.61
Lag	0.04	0.01	2.65[b]
Target Familiarity	−0.02	0.04	−0.50
Foil Familiarity	−0.05	0.03	−1.40
Target	0.52	0.17	3.05[b]

[a] Freq. = frequency, −2 log likelihood = 2462.42.
[b] $p < 0.01$.
[c] $p < 0.0001$.
[d] $p < 0.05$.

error made by pharmacists and college students. In addition to illustrating *that* similarity is associated with errors, the psychological model explains *why* similarity causes memory errors. Our data were supportive of Anderson's ACT-R model of recognition memory (&erson and Lebiere, 1998). The ACT-R model asserts that similarity increases the probability of false recognition errors because it increases the probability of noisy partial matching between new and old words in the recognition memory task. Although it was focused narrowly on the problem of drug name confusions, our study also makes a contribution to the growing literature on false memories (Roediger and McDermott, 1995). We demonstrated that both experts (pharmacists) and novices (college students) were vulnerable to false memory biases. Pharmacists made far fewer errors overall than their novice counterparts, a difference we attributed to familiarity with drug names. However, the positive effect of similarity on probability of false recognition was essentially unchanged by level of expertise. To our knowledge, this is the first investigation to examine the false memory phenomenon under different levels of expertise.

Limitations

The studies reported above were limited in several respects. First of all, we examined only clinical pharmacists and college undergraduates. Generalization of the observed effects to non-pharmacist health professionals or to the lay population at large may not be warranted. One should even be cautious generalizing the results to the broader population of community pharmacists since our participants were all clinical pharmacists from an academic medical center. Second, because of sampling error in the original survey, some of the prescribing frequency data which we used lacked precision, especially estimates below 2 million prescriptions per year (i.e., log frequency less than 6) (Woodwell, 1999). Third, experiments 1 and 3 included very few generic drug names. Thus, the effects we observed in those experiments can be generalized with greatest confidence only to other brand names. The recognition task we used was not a perfect model of what pharmacists and patients do. In reality, pharmacists and patients must select the correct drug from an array of *simultaneously presented* alternatives. In contrast, we asked participants to give an 'old' or 'new' response to each name *individually*. All stimulus presentations were visual, even though it is well-known that visual presentations activate phonological representations in memory (Marslen-Wislon, 1989). Future work should attempt to partial out the unique contributions to error of phonological and orthographic similarity. Finally, the absolute error rates reported above should not be over-interpreted. The increasing probability of error as a function of similarity was most salient. The absolute rates were a function of similarity as well as stimulus durations, list lengths, and other extraneous characteristics of our specific experimental task.

Conclusion

Excessive similarity between drug names significantly increases the likelihood of false recognition memory errors among practicing pharmacists and college students. In terms of cognitive psychology, errors result from a noisy partial match between orthographic and phonological features of the intended drug and its confusingly similar neighbor. From the point of view of industrial quality control, similarity is an assignable source of variation in the look-alike/sound-alike medication error rate. As such, steps should be taken to minimize similarity between new drug names and existing drug names. Where confusion between previously approved names is likely to occur, any of a variety of error minimization strategies, including computerized order entry, use of bar codes, automated dispensing, and non-alphabetical storage of drug products, should be used (Schiff & Rucker, 1998; Bates et al., 1998; Cohen, 1999).

Acknowledgements

This research was supported in part by the US Pharmacopeia's Fellowship Program, the Drug Information Association, the National Patient Safety Foundation, the Latiolais Leadership Fund, and the Campus Research Board of the University of Illinois at Chicago. The authors acknowledge the helpful assistance of John Anderson, Dan Boring, Bill Brewer, Mike Cohen, Gary Dell, Sanjay K. Gandhi, Robert Gibbons, Prahlad Gupta, Keith Johnson, David Lambert, Eric Lambert, Robert Lee, Don Rucker, Susan Schappert, Gordon Schiff, Tom Shaffer and William Wallace. Preliminary results of this research were reported at the conference on Enhancing Patient Safety and Reducing Errors in Health Care, November 8-10, 1998, Rancho Mirage, CA and at the Annual Meeting of the American Association of Pharmaceutical Scientists, November 15-19, 1998, San Francisco, CA. Requests for reprints should be sent to Dr. Lambert.

References

Allan, E. L., & Barker, K. N. (1990). Fundamentals of medication error research. *American Journal of Hospital Pharmacy*, *47*, 555–571.

Anderson, J. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, *128*, 186–197.

Anderson, J. R. (1995). *Learning and memory: An integrated approach*. New York: Wiley.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*, 341–380.

Anderson, J. R., & Bower, G. H. (1974). Interference in memory for multiple contexts. *Memory & Cognition*, *2*, 509–514.

Anderson, J. R., Lebiere, C. (Eds.) (1998). *The atomic components of thought*. New Jersey: Erlbaum, Mahwah.

Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, *104*, 728–748.

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, *30*, 221–256.

Anisfeld, M., & Knapp, M. (1968). Association, synonymity, and directionality in false recognition. *Journal of Experimental Psychology*, *77*, 171–179.

Anonymous (1999). Americans on track to fill record 3 billion prescriptions in 1999. [cited September 17, 1999]; Available from: URL: http://cnn.com/HEALTH/9908/29/AM-More-Drugs.ap/index.html.

Bates, D. W., Cullen, D. J., Laird, N., Petersen, L. A., Small, S. D., Servi, D., Laffel, G., Sweitzer, B. J., Shea, B. F., & Hallisey, R. (1995). Incidence of adverse drug events and potential adverse drug events. *Journal of the American Medical Association*, *274*, 29–34.

Bates, D. W., Leape, L. L., Cullen, D. J., Petersen, L. A., Teich, J. M., Burdick, E., Hickey, M., Kleefields, S., Shea, B., Vander Vliet, M., & Seger, D. L. (1998). Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. *Journal of the American Medical Association*, *280*, 1311–1316.

Bencomo, A. A., & Daniel, T. C. (1975). Recognition latency for pictures and words as a function of encoded-feature similarity. *Journal of Experimental Psychology: Human Learning and Memory*, *104*, 119–125.

Berwick, D. M. (1991). Controlling variation in health care: A consultation from Walter Shewhart. *Medical Care*, *92*, 1212–1225.

Berwick, D. M., & Leape, L. L. (1999). Reducing errors in medicine. *British Medical Journal*, *319*, 136–137.

Boring, D. (1997a). The development and adoption of nonproprietary, established, and proprietary names for pharmaceuticals. *Drug Information Journal*, *31*, 621–634.

Boring, D., Homonnay-Weikel, A. M., Cohen, M., & Di Domizio, G. (1996). Avoiding trademark trouble at FDA. *Pharmaceutical Executive*, *16*, 80–88.

Boring, D. L. (1997b). The CDER labeling and nomenclature committee: Structure, function, and process. *Drug Information Journal*, *31*, 7–11.

Boring, D. L., Stein, I. A., & Di Domizio, G. (1998). United States: Trademark trainwrecks at FDA. *Trademark World*, *108*, 29-33.

Cedrus (1991). SuperLab. [cited (1999 October 8, 1999]; Available from: URL: http://www.cedrus.com/.

Cohen, M. (Ed.) (1999). *Medication errors*. Washington, DC: American Pharmaceutical Association.

Davis, N. M. (1997). Drug names that look and sound alike. *Hospital Pharmacy*, *32*, 1558–1570.

Drewnowski, A., & Murdock Jr., B. B. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 319–332.

Hedeker, D. (1999). MIXOR/MIXREG Website. [cited 1999 June 4]; Available from: URL: http://www.uic.edu/~hedeker/mix.html.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.

Hedeker, D., & Gibbons, R. D. (1996). MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, *49*, 157–176.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Kleinbaum, D. G. (1994). *Logistic regression: a self-learning text*. New York: Springer.

Kohn, L. T., Corrigan, J., & Donaldson, M. S. (Eds.). (2000). *To err is human: building a safer health system*. Washington, DC: Institute of Medicine.

Lambert, B. (1997). Predicting look- and sound-alike medication errors. *American Journal of Health-System Pharmacy*, *54*, 1161–1171.

Lambert, B. L., Lin, S. -J., & Gandhi, S. K. (1997). Using normalized edit distance to assess the likelihood of look-alike and sound-alike medication errors. *Pharmaceutical Research*, *14*, S–233.

Lambert, B. L., Lin, S. -J., Chang, K. Y., & Gandhi, S. K. (1999). Similarity as a risk factor in drug name confusion errors: The look-alike (orthographic) and sound-alike (phonological) model. *Medical Care*, *37*, 1214–1225.

Lawrence, D. (1999). *Is medical care obsolete*? (pp. 1–14). July 14, 1999. Washington, DC National Press Club.

Leape, L. L., Bates, D. W., Cullen, D. J., Cooper, J., Demonaco, H. J., Gallivan, T., Hallisey, R., Ives, J., Laird, N., Laffel, G., Nemeskal, R., Petergen, L. A., Porter, K., Servi, D., Shea B. F., Small, S. D., Sheitzer, B. J., Thompson, B. T. (1995). Systems analysis of adverse drug events. *Journal of the American Medical Association*, *274*, 35–43.

Marslen-Wislon, W. (Ed.) (1989). *Lexical representation and process*. Cambridge, MA: MIT Press.

Nelson, D. L., & Davis, M. J. (1972). Transfer and false recognition based on phonetic identities of words. *Journal of Experimental Psychology*, *92*, 347–353.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction*. New York: Harcourt Brace.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1992). Models for recognition and recall. *Annual Review of Psychology*, *43*, 205–234.

Raser, G. (1972). False recognition as a function of encoding dimension and lag. *Journal of Experimental Psychology*, *93*, 333–337.

Reason, J. (1990). *Human error*. New York: Cambridge University Press.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814.

Schappert, S. M. (1994). National Ambulatory Medical Care Survey: 1992 summary. *US National Center for Health Statistics. Advance Data from Vital and Health Statistics*, *253*, 1–20.

Schappert, S. M. (1996). National Ambulatory Medical Care Survey: 1994 summary. *US National Center for Health Statistics. Advance Data from Vital and Health Statistics*, *273*, 1–18.

Schiff, G. D., & Rucker, T. D. (1998). Computerized prescribing: Building the electronic infrastructure for better medication usage. *Journal of the American Medical Association*, *279*, 1024–1029.

Selvin, S. (1996). *Statistical analysis of epidemiological data*. New York: Oxford University Press.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM — retrieving efficiently from memory. *Psychonomic Bulletin and Review*, *4*, 145–166.

Slack, M. (1991). Drug name confusion. *Lancet*, *338*, 190–191.

Sommers, M. S., & Lewis, B. P. (1999). Who really lives next door: Creating false memories with phonological neighbors. *Journal of Memory and Language*, *40*, 83–108.

Stephen, G. A. (1994). *String searching algorithms*. River Edge, NJ: World Scientific.

U. S. Department of Health and Human Services. (1996). 1993 National Ambulatory Medical Care Survey. [cited 1999 September 23, 1999]; Available from: URL: http://www.cdc.gov/nchswww/products/catalogs/subject/cdprice.htm#namcscdprice.

U. S. Pharmacopeia. (1993). *Drug product quality review*. March. Report No. 34. Rockville, MD: U. S. Pharmacopeia (pp.1–2).

U. S. Pharmacopeia. (1995). *USP quality review*. July. Report No. 49. Rockville, MD: U. S. Pharmacopeia.

U. S. Pharmacopeia. (1996). *USP quality review*. In *OTC names: An invitation to err*. (pp. 1–3) Rockville, MD: U. S. Pharmacopeia.

U. S. Pharmacopeia. (1997). Similar names. *USP DI Update vols. I and II*, 4–7.

U. S. Pharmacopeia. (1998). *USP dictionary of USAN and international drug names*. Rockville, MD: U. S. Pharmacopeia.

Underwood, B. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology*, *70*, 122–129.

Underwood, B. J. (1970). Word frequency and short-term recognition memory. *American Journal of Psychology*, *83*, 343–351.

Underwood, B. J., & Zimmerman, J. (1973). The syllable as a source of error in multisyllable word recognition. *Journal of Verbal Learning and Verbal Behavior*, *12*, 701–706.

Wallace, W. P. (1968). Incidental learning: The influence of associative similarity and formal similarity in producing false recognition. *Journal of Verbal Learning and Verbal Behavior*, *7*, 50–54.

Wallace, W. P., Stewart, M. T., & Malone, C. P. (1995a). Recognition memory errors produced by implicit activation of word candidates during processing of spoken words. *Journal of Memory and Language*, *34*, 417–439.

Wallace, W. P., Stewart, M. T., Shaffer, T. R., & Wilson, J. A. (1998). Are false recognitions influenced by prerecognition processing?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 299–315.

Wallace, W. P., Stewart, M. T., Sherman, H. L., Heather, L., & Mallor, M. D. (1995b). False positives in recognition memory produced by cohort activation. *Cognition*, *55*, 85–113.

Woodwell, D. A. (1999). National Ambulatory Medical Care Survey: 1997 summary. *US National Center for Health Statistics. Advance Data from Vital and Health Statistics, 305*, 1–26.

Woodwell, D. A., & Schappert, S. M. (1995). National Ambulatory Medical Care Survey: 1993 summary. *US National Center for Health Statistics. Advance Data from Vital and Health Statistics*, *270*, 1–20.