WHEN IS CATEGORY SPECIFIC IN ALZHEIMER'S DISEASE?

Keith R. Laws[1]*, Tim M. Gale[2], Verity C. Leeson[3] and John R. Crawford[4]

([1]Brain and Cognition Research Group, Psychology Division, Nottingham Trent University, UK; [2]Queen Elizabeth Hospital, Welwyn Garden City, UK; [3]Department of Psychology, London Metropolitan University, UK; [4]Department of Psychology, University of Aberdeen, UK)

ABSTRACT

Mixed findings have emerged concerning whether category-specific disorders occur in Alzheimer's disease. Factors that may contribute to these inconsistencies include: ceiling effects / skewed distributions for control data in some studies; differences in the severity of cognitive deficit in patients; and differences in the type of analysis (in particular, if and how controls are used to analyse single case data).We examined picture naming in Alzheimer's patients and matched elderly healthy normal controls in three experiments. These experiments used stimuli that did and did not produce ceiling effects / skewed data in controls. In Experiment 1, we examined for category effects in individual DAT patients using commonly used analyses for single cases (χ^2 and z-scores). The different techniques produced quite different outcomes. In Experiment 2a, we used the same techniques on a different group of patients with similar outcomes. Finally, in Experiment 2b, we examined the same patients but (a) used stimuli that did not produce ceiling effects / skewed distributions in healthy controls, and (b) used statistical methods that did not treat the control sample as a population. We found that ceiling effects in controls may markedly inflate the incidence of dissociations in which living things are differentially impaired and seriously underestimate dissociations in the opposite direction. In addition, methods that treat the control sample as a population led to inflation in the overall number of dissociations detected. These findings have implications for the reliability of category effects previously reported both in Alzheimer patients and in other pathologies. In particular, they suggest that the greater proportion of living than nonliving deficits reported in the literature may be an artifact of the methods used.

Key words: Semantic memory, dissociations, ceiling effects

INTRODUCTION

Patients with Alzheimer's disease have a well-documented picture naming impairment and some studies have further revealed the presence of category specific deficits. Nevertheless, the incidence and pattern of category-specificity across Alzheimer's patients *as a group* (Silveri et al., 1991; Tippett et al., 1996), and for *individual* Alzheimer's patients (Mauri et al., 1994; Garrard et al., 1998; Gonnerman et al., 1997; Laws et al., 2002; Laws et al., 2003) has been inconsistent. Most have reported living deficits, a minority has reported nonliving deficits, some report both and still others find no category specific effects at all in Alzheimer's patients. Several questions remain unanswered including: what factors might contribute to whether studies do or do not find category effects; whether a living or nonliving category effect is reported; and why so many living cases are reported?

First, category-specific effects may, of course, be hidden and/or distorted within a group analysis because individual Alzheimer's patients have category effects in opposing directions (i.e. some living and some nonliving) and so, cancel each other. In a cross-sectional design, Gonnerman et al., (1997) reported that the presence of living or nonliving deficits was related to the degree of anomia, i.e. patients with less impaired naming showed a deficit for nonliving things, and those more severely impaired showed living thing deficits. While this might explain some variability across previous studies, two recent studies have failed to replicate the reported association in larger samples of Alzheimer's patients (Zannino et al., 2002; Garrard et al., 1998). Nevertheless, the importance of examining individual patients and their heterogeneity is emphasized as a critical factor.

It is also notable that previous studies have examined Alzheimer's patients within a restricted range of cognitive ability as indicated by their MMSE scores (see Table I and below), and a constrained and *distorted* range of ability in controls (see Table II). Indeed, for a long time in this literature, it has been assumed that patient performance is an *exaggerated* version of the *assumed* normative profile. Although this hypothesis had not been explicitly examined, it was assumed that normal subjects would similarly find those items more difficult to name which are: less familiar, have lower name frequencies, greater visual complexity and so on - in other words, living things; and this was used to partly explain the 5:1 ratio for living to nonliving deficits. More recently, however, evidence has emerged that, counter to this assumption, it is quite common for normal subjects to show better naming of living than nonliving things (for examples using a variety of paradigms, see: Laws 1999, 2000, 2002; Laws and Gale 2002, 2002b; Laws and Neve 1999; Laws et al., 2002).

In this context, it is essential to also examine patients with a wider range of impairment than previously examined (especially with regard to notions of an interaction between category deficit and severity e.g. Gonnerman et al., 1997).

TABLE I
Levels of general cognitive functioning in studies of category specificity in Alzheimer's Patients

Study	General cognitive functioning
Silveri et al., 1991	described as 'mild'
Montanes et al., 1995	MMSE = 17-26
Tippet et al., 1996	described 9 as mild and 5 as moderate
Gonnerman et al.,1997	MMSE = 19 and 18
Garrard et al., 1998	MMSE = 19.9
Garrard et al., 2001	MMSE = 23.6
Zannino et al., 2002	MMSE = 20.6

TABLE II
Living and nonliving naming levels for control subjects in studies of category specificity in Alzheimer's Patients

Study	Living Mean (%)	Nonliving Mean (%)
Silveri et al., 1991	99.8	99.8
Mauri et al., 1994	98	98
Montanes et al., 1995	96	97
Gonnerman et al., 1997	97	97.2
Garrard et al., 1998	93	97
Garrard et al., 2001	90.5	93.3
Zannino et al., 2002	98	98.3

Similarly, in those studies that have included normal controls, their naming is invariably at or near ceiling (see Table II). These control ceiling effects probably reflect the widespread reliance upon simple line drawn stimuli that controls have little difficulty in naming under normal conditions (the Snodgrass and Vanderwart corpus is used in 90+% of category specific studies: Laws and Gale 2002a,b). Studies comparing patient performance with that of ceiling level performance in controls will distort the findings for patients (including both the degree and possibly the types of deficit reported: see Laws et al., 2002, 2003; Fung et al 2001).

Finally, the literature displays a surprising lack of agreement about the necessary and sufficient conditions for documenting a category specific naming disorder (see Laws 1998 for a discussion). Indeed, a variety of methods for delineating category effects have been used in studies of Alzheimer's patients alone. These include: (a) within-patient comparison of living and nonliving naming, but without control data (e.g., Tippett et al., 1996); (b) within-patient comparison of living and nonliving naming without control data or statistical analysis (e.g. Gonnerman et al., 1997); (c) comparisons of the raw difference scores (i.e. living minus nonliving) for patients against control naming cut-offs from a non-matched sample (e.g. Garrard et al 1998); (d) the comparison of individual patient scores for living and nonliving separately using *z*-scores derived from control data (e.g. Mauri et al 1994; for problems with this approach, see Laws et al., 2003); and (e) regression analyses where control baseline data are included (Zannino et al., 2002 Laiacona et al., 1998). There is, therefore, no consensus on how to actually document, and hence define, a category specific disorder; this is as true of studies with Alzheimer's patients as those with other pathologies (e.g. herpes simplex encephalitis).

The current study examines the implications of these issues and proposes an accessible and readily usable method to rectify some of the problems. Experiment 1 examines the consequences of determining category effects when using typical approaches that either include healthy controls (e.g. *z*-scores, Z_D) or are limited to within-patient comparisons (e.g. χ^2). Experiment 2a and Experiment 2b compares naming in a new series of Alzheimer's patients with that of healthy controls who are either performing at ceiling or not.

Finally and importantly, gender affects both category naming and category fluency in healthy controls (see Laws 1999, 2000, 2003) and Alzheimer's patients (Laiacona et al., 1998), where appropriate, all comparisons were with healthy controls who were matched both for age and for sex i.e. all *z* and *t*-values were referenced from gender specific comparisons.

Methods

Experiments 1 and 2a use methods that are typically applied to determine the presence of category effects. The stimuli (from the Snodgrass and Vanderwart corpus 1980) are those typically used in category effect studies (Laws and Gale 2002a,b); and the analyses used cover the majority of approaches to documenting category effects for individual patients whatever the pathology. We evaluate the difference outcomes associated with using and not using control data in determining category effects (by comparing the outcomes for χ^2 versus *z*-score analyses); and the subsequent limitations of these approaches with these type of data.

TABLE III

Naming raw scores (and percentages) for Alzheimer's patients across category, along with significant χ2 and Z-score analyses (Experiment 1)

Patient	Living	Nonliving	χ^2	Z[a]	Z_D[b]
A1	5 (25%)	7 (37%)	–	L + NL	L
A2	15 (75%)	14 (74%)	–	L	L
A3	15 (75%)	9 (47%)	–	L + NL	L
A4	11 (55%)	12 (63%)	–	L	L
A5	6 (30%)	6 (32%)	–	L + NL	L
A6	12 (60%)	9 (47%)	–	L + NL	L
A7	18 (90%)	10 (53%)	NL	L + NL	–
A8	19 (95%)	15 (79%)	–	–	–
A9	16 (80%)	12 (63%)	–	L + NL	L

[a] One-tailed. Eighteen comparisons adjusted using Bonferroni (.05/18 = .0027 i.e. Z > 2.78)
[b] Two-tailed. Nine comparisons adjusted using Bonferroni correction

The types of statistical analyses also differ widely across studies, ranging from parametric group comparisons (which are likely to be inappropriate for example when controls perform at ceiling as happens in many category effect studies). At an individual case level, studies use normal control data e.g. *z*-scores or *z*-score differences (Z_D) and within-patient χ^2 analyses (being used in approximately 70% of category-specific studies). In the case of Z_D, the method outlined by Payne and Jones (1957) and described in Crawford et al. (1998) was used. This method divides the difference between a patient's *z* scores by the standard deviation of the difference in the controls (s_{X-Y}) to obtain a *z* score for the difference (Z_D). The Z_D can then be referred to a table of the areas under the normal curve to test whether it exceeds the required critical value (i.e., 1.96 for a two-tailed test). The standard deviation of the difference is

$$s_{X-Y} = \sqrt{2 - 2r_{XY}}, \quad (1)$$

where r_{XY} is the correlation between the two tests in the control sample used to obtain the patient's *z*-scores, and the first value under the square root sign (2) is the sum of the *SD*s for the two tests in the control sample (*Z* scores have *SD*s of 1).

Experiment 1

Subjects

(a) Alzheimer's patients (n = 9)

Nine patients (7 female; 2 male: mean age = 81.1 years) diagnosed with probable Alzheimer's disease according to NINCDS-ADRDA criteria (McKhann et al., 1984). The patients had a mean MMSE (Folstein et al., 1975) of 13.7 ± 3.74

(b) Normal healthy elderly participants (n = 12)

Twelve (8 female; 4 male) elderly control subjects (mean age = 77.9 years), were recruited from drop-in centres for the elderly (being residents in nursing homes or visitors to community day-centres). All were screened for good health, and had no history of head injury, neurological or psychiatric illness, nor alcohol or drug abuse.

Materials and procedure

Forty line drawings (depicting 20 living and 20 non-living things: see Appendix 1) were selected from the Snodgrass and Vanderwart (1980) corpus. The two sets were matched for familiarity (3.12 vs. 3.16, p = .85); visual complexity (3.26 vs. 3.25, p = .95); and name frequency (10.95 vs. 9.55, p = .7).

Results

The Alzheimer's group correctly named 60 ± 19% (60.11 ± 19.28% living and 55 ± 16.97% nonliving) of the stimuli, while the control group named 95 ± 6% (94.66 ± 6.13% living and 90.35 ± 9.74% nonliving)[1]. The difference between living and nonliving naming scores did not correlate significantly with overall naming ability (r = .5, p = .17) or with MMSE (r = .18, p = .64) scores. Nevertheless, these correlations reflected a narrow range of quite low MMSE scores in nine patients.

Individual cases

The use of within-patient analysis, using a chi square test of independence (χ^2), produced only one deficit (for nonliving things). By contrast, *z* scores showed that all patients except one (A8) were impaired; all for living things and all except two (A2 and A4) were impaired for nonliving things (see Table III for comparisons). However, as the use of χ^2 is essentially aimed at detecting dissociations (rather than simple deficits) between living and nonliving things, the most relevant comparison is

[1] The item 'flute' was removed from the analysis because it elicited errors in the majority of control group subjects.

with the results of the Payne and Jones test (Z_D). Here again, however, the use of within-patient comparisons (χ^2) versus Z_D (which involves referencing the patient's performance to controls) yields conclusions that could scarcely be more different; the Payne and Jones test revealed that all except two (A7 and A8) patients showed significantly worse living than nonliving naming. In contrast, and as noted, using χ^2 only one patient with a dissociation was detected and this was for nonliving things.

Summary

The main outcomes of this experiment concern: the different incidence of dissociations reported when using within (χ^2) versus between-subject (Z_D) analyses; and the lack of correspondence between the direction of absolute score differences and the direction indicated when patient performance is referenced to normative data.

Although very common in the case study literature, χ^2 revealed only one deficit in patient A7 (for nonliving things), who showed the largest absolute difference of any patient. However, ironically given the large number of dissociations detected using the Payne and Jones test in the present sample, patient A7 did not exhibit a significant dissociation (i.e. Z_D was < 1.96 for this patient). Therefore, although the direction and magnitude of living-nonliving differences in patients may look large, they need not be *abnormal*. This patently shows that absolute raw differences (i.e. when unreferenced to control data) may misrepresent both the number and the direction of category deficits reported. A7 would be an interesting case because, some researchers might consider worse nonliving than living naming to run counter to the expected pattern (predicted by familiarity, name frequency and visual complexity etc) and so, could view A7 as a clear case of a nonliving deficit.

Experiment 2a

Experiment 1 showed how it is necessary to reference patient data to that of controls otherwise we are likely to make quite profound errors in determining category effects. Experiment 2a uses the same methods outlined in Experiment 1 to examine category effects in a larger group of less severely ill Alzheimer's patients.

Subjects

(a) Alzheimer's patients (n =18)

Eighteen patients (14 female; 4 male: mean age = 77.57 ± 8.22: 77.25 ± 4.79) with probable Alzheimer's dementia according to NINCDS-ADRDA criteria (McKhann et al., 1984) were tested. Their mean MMSE was 18.03 ± 4.69 (cf. mean MMSE = 13.7 ± 3.74 for patients in Experiment 1). All patients were living at home and visiting a day-center.

(b) Normal healthy elderly participants (n = 22)

Twenty-six normal subjects (10 female; 12 male: mean age = 71.9 ± 4.33: 71.17 ± 3.79) were recruited through their general practitioner, who screened them for good health. They had no history of head injury, neurological or psychiatric illness, or alcohol or drug abuse. English was the first language for all participants.

Materials and procedure

Sixty-four line drawings (see Appendix 1) were taken from the Snodgrass and Vanderwart corpus (1980). The pictures comprised 32 living and 32 nonliving items that were matched across category for familiarity (2.82 vs. 2.97, p = .39; visual complexity (3.21 vs. 3.09, p = .47) and name frequency (9.66 vs. 13.23, p = .3).

Results

As expected, naming was extremely impaired in the Alzheimer's patients (66.11 ± 19.8% [65.62 ± 20.31 living and 66.59 ± 21.56 nonliving]) and the mean naming score for the healthy controls was at ceiling (95.67 ± 3.97% [95.91 ± 4.31 living and 95.44 ± 7.19% nonliving]). Analysis across category at the group level revealed no difference between living and nonliving naming for the Alzheimer's patients. The correlation between the living-nonliving difference and overall naming was non-significant (r = –.14, p = .58). MMSE scores also again failed to correlate with the difference score (r = – .04, p = .87).

Individual cases

Again χ^2 revealed a low incidence of deficits (2/18: 11%: AF and AH both living deficits) compared to Z_D (6/18: 33%: which included AF and AH). Also again, the *raw* score for one patient (AR) was lower for nonliving than living but AR nevertheless showed a dissociation in the opposite direction when scores were referred to control values using Z_D (see Table IV).

Summary

The proportion of deficits was again very different when comparing the within-patient method with the use of controls (2 versus 6). The incidence of dissociations is much lower than that of the patients in Experiment 1, but this is wholly

TABLE IV

Naming raw scores (and percentages) for Alzheimer's patients across category, along with significant χ^2 and Z-score analyses (Experiment 2a)

Patient	Living	Nonliving	χ^2	Z[a]	Z_D[b]
AA	31 (97%)	30 (94%)	–	–	–
AB	29 (91%)	31 (97%)	–	–	–
AC	29 (91%)	26 (81%)	–	–	–
AD	29 (91%)	25 (78%)	–	–	–
AE	25 (78%)	29 (91%)	–	L	–
AF	18 (56%)	28 (88%)	L	L	L
AG	23 (72%)	23 (72%)	–	L + NL	–
AH	18 (56%)	27 (84%)	L	L	L
AI	21 (66%)	23 (72%)	–	L + NL	–
AJ	24 (75%)	19 (59%)	–	L + NL	–
AK	19 (59%)	22 (69%)	–	L + NL	L
AL	21 (66%)	19 (59%)	–	L + NL	–
AM	19 (59%)	20 (63%)	–	L + NL	–
AN	19 (59%)	17 (53%)	–	L + NL	–
AO	16 (50%)	17 (53%)	–	L + NL	L
AP	14 (44%)	11 (34%)	–	L + NL	–
AQ	11 (34%)	13 (41%)	–	L + NL	L
AR	12 (38%)	9 (28%)	–	L + NL	L

[a] One-tailed. Thirty-six comparisons adjusted using Bonferroni (.05/36 = .001)
[b] Two-tailed. Eighteen comparisons adjusted using Bonferroni correction

consistent with the lower MMSE scores of that group. Again, as for Experiment 1, in 2a the incidence of dissociations documented with χ^2 is much lower than with Z_D (being 1:7 and 1:3 respectively). Given that Experiment 1 contained patients who are more severely ill, this suggests that the ratio of living: nonliving dissociations will be more than twice that of less impaired patients. Experiment 2a documents only dissociations involving differential impairment for living things. The finding of the far greater incidence of living thing cases is consistent with the published literature (the ratio of living: nonliving being approximately 5:1 across all pathologies).

The major finding from Experiments 1 and 2a, however, relates not to the performance of the patients, but to that of the control subjects. The ceiling performance of controls will undoubtedly influence outcomes and makes some analyses untenable (they are especially likely to distort *z*-scores particularly when this is combined with small sample sizes). Therefore, as a direct comparison, in Experiment 2b we test the same patients as Experiment 2a but (a) using stimuli that do not produce ceiling effects in controls and (b) we also report results from inferential methods that are more appropriate than the use of *z*-scores for detecting deficits and dissociations.

Experiment 2b

We have shown in Experiment 1 and 2a that control data are essential, and that simple within-patient comparisons are misleading. Additionally, the use of *z*-scores with small normative samples will overestimate the degree of impairment (and inflate the Type I error rate) in patients because the statistics of the control sample are treated as population parameters rather than sample statistics (see Crawford and Garthwaite, 2002). Hence, new methods were used to analyse the data in Experiment 2b (outlined below).

Previous studies comparing naming of Alzheimer's patients and normal controls have largely ignored the fact that normal naming is typically at, or near, ceiling on simple naming tasks. On the widely used Snodgrass and Vanderwart corpus (1980) of line drawings, the range has been very high (93-99%: see introduction). Indeed, the controls both in Experiment 1 (95 ± 6%) and in Experiment 2a (95.6 ± 3.97%) performed at ceiling. The statistical methods employed in Experiments 1 and 2a compare with those employed in past studies, but will distort the pattern of findings (because of the ceiling effect in controls and resultant skewed distribution of scores). In this experiment, therefore, we compared naming of the same patients and controls as in Experiment 2, but on 60 living and 60 nonliving pictures that are graded in difficulty. The purpose being both to avoid a ceiling effect in controls and to compare consistency across stimulus sets within the same patients.

Method

The data for each individual subject were examined by comparing performance with controls using methods described by Crawford and Garthwaite (2002) for testing for deficits and dissociations in single-case studies; see also Crawford, Garthwaite and Gray (2003)[2]. Initially, patient naming was

[2] An alternative method using parametric and nonparametric tolerance limits has been proposed by Capitani and colleagues (see Capitani et al., 1999). Unfortunately, this method relies upon testing very large normative samples (between 400 and 1000: see Wald 1943; and Mycroft et al., 2002 for a further recent discussion of this method) and so (for pragmatic reasons) has rarely been used in the general cognitive neuropsychological literature. In the case of the current study, our numbers of controls are commensurable or larger than in most similar studies

compared with that of age- and gender-matched controls for living and nonliving things separately. This analysis (Crawford and Howell, 1998; Crawford and Garthwaite, 2002) determines whether an individual's score is significantly different from a control or normative sample and provides a point estimate of the abnormality of the scores -i.e. it estimates the percentage of the population that would obtain a lower score. The formula for this method, which is essentially a modified independent samples *t*-test, is as follows:

$$t = \frac{X - \bar{X}}{s\sqrt{\frac{N+1}{N}}}, \qquad (2)$$

where *s* is the control sample standard deviation; X is the patient's score, $\bar{X}$ is the mean score for controls, and N is the size of the normative sample.

Of course, it is possible for patients to be impaired at naming living or nonliving things, but that the *difference* between their scores does not reach significance; conversely, a patient may be severely impaired on both tasks but still show differential impairment. Therefore, for those patients showing impaired naming of living and/or nonliving things, we compared their living-nonliving discrepancy score with the mean discrepancy of the normative sample (Crawford et al., 1998; Crawford and Garthwaite, 2002). This method tests whether the discrepancy observed for the patient is significantly different from the discrepancies observed for controls and provides a point estimate of the abnormality of the individual's discrepancy -i.e. it estimates the percentage of the population that would obtain a more extreme discrepancy. The formula for this method, which is a modified paired samples *t*-test, is as follows:

$$t = \frac{|X - \bar{X}|}{\sqrt{(2 - 2r_{XY})\left(\frac{N+1}{N}\right)}}, \qquad (3)$$

where Z_X and Z_Y are the individual's scores on tests X and Y expressed as *z*-scores based on the means and *SD*s of the controls, and r_{XY} is the correlation between the test scores for the control sample. These methods of testing for deficits and for differences (i.e. dissociations) are to be preferred over the use of *z* and Z_D as they treat the statistics of the control sample as statistics rather than as population parameters[3].

Crawford et al. (2003) have recently provided fully specified criteria for Shallice's (1988) classification of 'strong' and 'classical' dissociations and we employ these in the present study. A patient was considered to exhibit a *strong* dissociation if they were (a) impaired at naming both living *and* nonliving on Crawford and Howell's test (one-tailed) and (b) showed a significant difference (two-tailed) between the two scores on Crawford et al's (1998) test. A patient was considered to exhibit a classical dissociation if they were impaired at only living or nonliving naming (one-tailed) and showed a significant difference (two-tailed) between the impaired and intact category. Monte Carlo simulations indicated that using these criteria, the probability of incorrectly classifying an individual drawn from the healthy population as exhibiting either of these types of dissociation is low (Crawford et al., 2003).

Subjects

The same patients and controls tested in Experiment 2a.

Materials and procedure

Participants viewed 120 colour pictures from the Category Specific Names Test (McKenna 1997; McKenna and Parry 1994); this included 60 living and 60 nonliving things that are matched across category for familiarity (3.31 vs. 3.23, $p = .50$) and name frequency (Baayen et al 1993: 37.3 vs. 32.5, $p = .67$). The 30 pictures in each category were presented in order of normative naming-difficulty. A list of items can be found in Appendix 1.

Results

Skewness and kurtosis statistics (g_1 and g_2) were computed for the male and female healthy control data. Skewness for males for living stimuli was –0.85 and was 0.24 for nonliving stimuli. D'Agostino et al. (1990) test for skewness failed to reject the null hypothesis that the distributions were symmetrical; $z_{g_1} = 1.36$, $p = 0.177$ for living, and $z_{g_1} = 0.39$, $p = 0.69$ for nonliving. Further, D'Agostino's omnibus test for normality, which uses both g_1 and g_2 as input, revealed that the distributions did not differ significantly from normality; $K^2 = 2.18$, $p = 0.34$ for living things, and $K^2 = 2.18$, $p = 0.34$ for living things.

In female controls, skewness for living stimuli was – 0.22 and was – 0.61 for nonliving stimuli. The distributions did not contain significant asymmetry; $z_{g_1} = 0.33$, $p = 0.74$ for living, and $z_{g_1} = -0.91$, $p = 0.36$ for nonliving. Also, as was the case for the male controls, the distributions did not depart significantly from normality using the omnibus test; $K^2 = 1.65$, $p = 0.44$ for living things, and $K^2 = 1.64$, $p = 0.44$ for living things. Thus, unlike the stimulus sets used in experiments 1 and

[3] Programs to run these analyses can be downloaded from http://www.abdn.ac.uk/~psy086/dept/psychom.htm

2a (and those used in previous studies), this demonstrates that it is clearly possible to obtain stimulus sets that yield normally distributed scores in controls.

Naming was again extremely impaired in these Alzheimer's patients (23.1%). As expected, the mean naming score for healthy controls was below ceiling (73.1%). Analysis across category at the group level revealed no difference between living and nonliving naming for either the Alzheimer's patients (14.78 ± 10.25 vs. 12.94 ± 9.65) or healthy controls (44.57 ± 7.2 vs. 44.92 ± 5.03). Again contrary to Gonnerman et al (1997) the correlation between the living-nonliving difference and naming ability failed to reach significance (r = –.11, p = .67) as did the correlation of MMSE scores with the difference score (r = .05, p = .85).

Individual cases

As in the previous experiments, analyses involving the comparison of individual patients with control data were conducted using gender matched controls cases. The mean score of female controls on living things was 46.20 (*SD* = 5.85) and was 45.50 (*SD* = 3.37) for nonliving things; the correlation between living and nonliving scores was 0.597. The mean score for male controls was 46.50 (*SD* = 6.91) on living things and was 46.75 (*SD* = 4.05) for nonliving things; the correlation between living and nonliving scores was 0.262.

As in the previous experiments, χ^2 produced a low incidence of deficits (3/18). By contrast, almost all patients were impaired on *z*-scores and single *t*-test analyses of living and nonliving naming separately. On the Z_D analysis, 14 patients showed significantly greater impairment (i.e. dissociations) for nonliving versus living things. This number was higher than the equivalent results obtained from the modified *t*-test: using this latter method, 10 patients exhibited significantly greater impairment for nonliving things (see Table V). The higher incidence obtained from the use of Z_D reflects an inflation of the Type I error rate (because Z_D inappropriately treats the control sample as a population).

Summary

The most striking finding from Experiment 2b was the incidence of differential deficits for *non*living things i.e. all 10/18 impaired patients were significantly more impaired for nonliving things. This contrasts strongly with the outcomes in Experiments 1 and 2a; and consequently with the prevailing 5:1 ratio of living: nonliving deficits recorded in the literature.

The second notable feature concerns the fact that some methods managed to produce a patient with a paradpxoc deficit across stimuli (for *z*-scores see AE and for Z_D see AQ). Indeed, all the living cases found in Experiment 2a either disappeared or became nonliving cases in Experiment 2b (see Table VI which for convenience combines the results of Experiment 2a and 2b).

Furthermore, the incidence of different category deficits may relate to the stimuli used – i.e. Experiment 2a produced 6 living thing deficits and Experiment 2b produced none; however Experiment 2b produced twice as many nonliving deficits as Experiment 2a. Hence the use of the Snodgrass and Vanderwart stimuli and the accompanying ceiling naming for controls may inflate the number of living thing deficits.

As with Experiment 1 and 2a, no evidence for a double dissociation (either *classical* or *strong*)

TABLE V

Naming raw scores (and percentages) for Alzheimer's patients across category, along with significant χ^2 and z-score analyses (Experiment 2b)

Patient	Living	Nonliving	χ^2	Z^a	Z_D^b	Modified independent *t*-test[a]	Modified paired *t*-test[b]
AA	34 (57%)	27 (45%)	-	NL	-	NL	-
AB	33 (55%)	34 (57%)	-	NL	-	-	-
AC	23 (38%)	14 (23%)	-	L+NL	NL	NL	*Strong NL*
AD	21 (35%)	7 (12%)	NL	L+NL	NL	L+NL	*Strong NL*
AE	31 (52%)	26 (43%)	-	NL	NL	NL	-
AF	10 (17%)	23 (38%)	L	L+NL	-	L+NL	-
AG	11 (18%)	7 (12%)	-	L+NL	NL	L+NL	*Strong NL*
AH	14 (23%)	18 (30%)	-	L+NL	-	L+NL	-
AI	5 (8%)	4 (7%)	-	L+NL	NL	L+NL	*Strong NL*
AJ	22 (37%)	16 (27%)	-	L+NL	NL	NL	*Strong NL*
AK	12 (20%)	10 (17%)	-	L+NL	NL	L+NL	*Strong NL*
AL	13 (22%)	14 (23%)	-	L+NL	NL	L+NL	-
AM	12 (20%)	15 (25%)	-	L+NL	NL	L+NL	-
AN	10 (17%)	7 (12%)	-	L+NL	NL	L+NL	*Strong NL*
AO	3 (5%)	6 (10%)	-	L+NL	NL	L+NL	*Strong NL*
AP	3 (5%)	2 (3%)	-	L+NL	NL	L+NL	*Strong NL*
AQ	4 (7%)	0 (0%)	NL	L+NL	NL	L+NL	*Strong NL*
AR	5 (8%)	3 (5%)	-	L+NL	NL	L+NL	-

[a] One-tailed. Thirty-six comparisons adjusted using Bonferroni (.05/36 = .001)
[b] Two-tailed. Eighteen comparisons adjusted using Bonferroni correction

TABLE VI
Category deficits in Alzheimer's patients (Comparing Experiment 2a and 2b)

	Experiment 2a			Experiment 2b				
	Sex	χ^2	*Z*	Z_D	χ^2	*Z*	Z_D^a	Dissociation
AA	M	–	–	–	–	NL	–	–
AB	M	–	–	–	–	–	–	–
AC	F	–	–	–	–	NL	NL	*Strong NL*
AD	F	–	–	–	NL	NL	NL	*Strong NL*
AE	F	–	L	–	–	NL	NL	–
AF	M	L	L	L	L	L + NL	–	–
AG	F	–	L + NL	–	–	L + NL	NL	*Strong NL*
AH	F	L	L	L	–	L + NL	–	–
AI	F	–	L + NL	–	–	L + NL	NL	*Strong NL*
AJ	F	–	L + NL	–	–	NL	NL	*Strong NL*
AK	F	–	L + NL	L	–	L + NL	NL	*Strong NL*
AL	F	–	L + NL	–	–	L + NL	NL	–
AM	F	–	L + NL	–	–	L + NL	NL	–
AN	F	–	L + NL	–	–	L + NL	NL	*Strong NL*
AO	F	–	L + NL	L	–	L + NL	NL	*Strong NL*
AP	F	–	L + NL	–	–	L + NL	NL	*Strong NL*
AQ	F	–	L + NL	L	NL	L + NL	NL	*Strong NL*
AR	M	–	L + NL	L	–	L + NL	NL	–

a Two-tailed. Eighteen comparisons adjusted using Bonferroni correction

emerged in Experiment 2b. It is notable, however, that the one method that did produce a double dissociation was the within-patient (χ^2) method (see AD and AQ versus AF [nonliving versus living]; though AQ was at floor almost). AD and AF do provide evidence of a double dissociation as it is often defined in the category specific literature. It is also clear that AF's living deficit disappears when referenced to control data. While AF showed a living deficit in Experiment 2a using all methods, AQ showed a contradictory deficit on Experiment 2a and AD showed none at all on Experiment 2a.

Finally, Experiment 2b indicates that the incidence of living and nonliving thing deficits may have been exaggerated and underestimated respectively. Thus, providing a possible explanation for the disproportionate number of living compared to nonliving deficits reported in this literature.

Discussion

These experiments highlight the importance of methodological issues concerning how we measure and define *category specific effects*. In particular, issues relating to (a) how the failure to use control subjects will distort the outcomes, creating Type I and II errors; and (b) that when controls are included, the typical performance of controls on the typical test materials will distort the deficit incidence across the two categories. By contrast, when patients are examined against controls whose naming is below ceiling, a quite different profile emerges for patients and in particular a greater incidence of nonliving deficits.

As already noted, studies of Alzheimer's patients have varied according to whether they used a control group and how the control data was used in analyses. Examination of the living-nonliving difference score (without reference to normal control data) does occur in studies of category effects in Alzheimer's patients (e.g. Gonnerman et al., 1997) and is if anything, the *norm* in studies of category-specific disorders resulting from other pathologies (Laws, in press). The current study, however, confirms that individual difference scores – if not considered in the context of control data – will produce both Type I (false positives) and Type II errors (false negatives). For example, Experiment 1 showed that the largest difference (A7: 90% living vs. 53% nonliving) was not significant when referenced to the naming of normal subjects (a potential false positive). By contrast, another patient (A5) showing no absolute difference in living and nonliving naming (30% vs. 31%) displayed a significant category deficit for living things (a potential false negative). Hence, the absolute size of difference between the ability of patients to name living and nonliving things will be misleading (be it exceptionally large or small) *unless* referenced to the normal naming pattern for that specific stimulus set. This is critical because the *absolute* difference in numbers of living and nonliving things named is very frequently used to define category-naming deficits in the category specific literature (and verified using χ^2). Laws (in press) found that this was the common practice in more than 80% of case studies examining category-specific deficits.

Nevertheless, all three experiments demonstrate how that the level of performance in the controls is critical. Examining patients of different severity and using methods and materials that are quite typical of this literature, Experiments 1 and 2a show that the concomitant ceiling effects in controls distort the incidence of category effects. In particular, both Experiment 1 and 2a indicate that when patient performance is referenced to control data that has a

ceiling effect, this dramatically increases the incidence of category effects (compared to within-patient comparisons of living and nonliving naming). Indeed χ^2 produces a small deficit incidence when compared to Z_D referenced deficits levels (Experiment 1: 1 vs. 7; Experiment 2a: 2 vs. 6).

Given the different level of abilities of the patients in Experiment 1 and 2a, the findings are not restricted to patients with a specific level of cognitive functioning (the level of cognitive functioning did not relate to the category of deficits, just the proportion reported i.e. being greater in more severely impaired patients). Within-patient χ^2 produced few deficits (3 in 27 patients), while Z_D produced (13/27 deficits). Moreover, the overwhelming majority of differential deficits (using both methods) were for living things (15/16 deficits) in Experiments 1 and 2a. By contrast, Experiment 2b (with the same patients and controls as 2a, but without the ceiling effect) exclusively documented differential deficits for nonliving naming. Hence it seems that highly accurate (i.e. ceiling or near-ceiling) control performance will distort the incidence of category effects documented insofar as it increases the number of living deficits and underestimates the incidence of nonliving deficits. Certainly our data suggest that the widespread use of simple line drawings – when referenced to ceiling effects in controls – will exaggerate the presence of living thing deficits. Given that the same stimuli and types of analysis are widely used in 90+% of studies of category specificity within other patient groups (see Laws and Gale 2002a, b; Laws in press), these limitations may be quite widespread in the literature.

It should be stressed that the approach taken in Experiment 2b was aimed at dealing with two distinct problems; the problem of skewed data /ceiling level performance in controls and the use of methods (i.e. *z* and Z_D) that treat the control sample a population. The present use of modified *t*-tests to compare individual patients with controls was used to address the second of these problems; it would be just as inappropriate to use modified *t*-tests with obviously non-normal data as it is to use methods based on *z*-scores (Crawford and Garthwaite, 2002; Crawford et al., 2003). Hence, the statistical methods advocated here are not proposed as remedies for the first problem referred to above; rather the remedy is to employ stimuli that do not produce a ceiling effect and skewed distributions in the control sample (Laws et al., 2003).

Furthermore, although the analyses in Experiment 2b highlighted how intra-individual methods of analysis (i.e. a χ^2 test on a patient's living and nonliving raw scores) can lead to erroneous conclusions, readers may be concerned that erroneous conclusions could also follow from application of the methods we advocate. In particular, it may be of concern that relatively small differences in raw scores (and even, as in Case AO, raw scores that are higher on the task recorded as more impaired) can lead to significant results and a patient being classified as exhibiting a dissociation. The living raw score for Case AO was 3 and the nonliving score was 6 and yet this patient fulfilled the criterion for a strong nonliving dissociation (i.e., performance on *non*living naming was more impaired than living). Part of such a concern is misplaced because the raw scores are misleading. The raw score for Case AO on living things when expressed as a *z* score based on the mean and *SD* of female controls was – 7.39 whereas the nonliving *z* score was – 11.72 (the nonliving score is more extreme because the control *SD* for nonliving items was smaller than that for living things).

It does, however, remain true that when a patient's performance on *both* tasks is *very* extreme (i.e. near floor: as with Case AO and three other cases in the present study i.e., AI, AP and AQ), the results of any classification method for dissociations should be treated with caution. That is, although the statistical tests may indicate a dissociation, the heuristic validity of such evidence is debatable. An assertion about differential nonliving deficits would, of course, be less compelling if based solely on cases such as the four described above. Nevertheless, differential deficits in nonliving naming were also observed for six other cases in whom performance was well above floor; as such, the pattern recorded in the extreme cases is consistent with the pattern observed in less impaired cases.

In applying Crawford et al.'s (2003) criteria for dissociations, we used three *t*-tests to infer the presence of a dissociation (tests were performed on a patient's living and non-living scores separately and a further test was applied to test whether the difference between the living and non-living scores was significant). An anonymous referee suggested that a potentially more parsimonious approach would be to apply a single test that made use of Hotelling's T^2 distribution. However, such a test would tell us only whether, overall, the patient differed from controls, but not if there was a dissociation (Hotelling's test would find the weighted composite of the two tasks that achieves the optimal discrimination between patient and controls) That is, a patient who performed very poorly but equivalently relative to controls on living and non-living naming would yield a significant result on Hotelling's test[4].

The main conclusions of the present study are that the consistency and reliability of category effects in Alzheimer's patients (and by implication,

[4] The suggestion of using Hotelling's T^2 distribution in comparing an individual patient's performance against controls is, however, an interesting one. Several potentially useful applications can be envisaged. For example, Crawford and Garthwaite (2003) have developed methods for testing whether the slope of patient's regression line is significantly different from a control sample. This work could be extended to test whether the combination of the slope of the patient's regression line and its intercept differed from controls. Hotelling's T^2 would be ideally suited to this purpose i.e. to providing a test on whether, overall, the patient's regression line was abnormal.

those with other pathologies) – are strongly influenced by: (a) the presence or absence of a control group; and (b) the presence or absence of ceiling effects /skew in the control data, and (c) the inferential method used to compare each patient with controls. These findings cast doubt upon the reliability of some previously reported category specific cases (at least as far as naming is concerned) in Alzheimer patients and in other pathologies. In this context, we would suggest that the findings from Experiment 2b are more likely to reflect something approaching a reliable pattern of results since (a) controls were used rather than drawing inferences from within-patient analysis of raw scores, (b) the controls performed comparably across categories *and* well below ceiling (≅ 75%), and (c) the inferential methods used to compare patient with controls are to be preferred over the use of *z* and Z_D for reasons outlined earlier.

Acknowledgements. We would like to thank the patients and controls for their participation; Prof. Chris Hawley, Dr Hazel Wood, and Dr Ann Wills for providing access to their patients; Dr Lia Kvavilashvili for helping with access to older healthy subjects; and three anonymous reviewers for their thoughtful and helpful comments on an earlier draft.

References

BAAYEN RH, PIEPENBROCK R, and VAN RIJN H. The CELEX lexical database. Philadelphia: Linguistic Data Consortium, University of Pennsylvania, 1993.

CAPITANI E, LAIACONA M, BARBAROTTO R and COSS F. How can interference be evaluated in attentional tests? A study based on bivariate nonparamteric tolerance limits. *Journal of Clinical and Experimental Neuropsychology, 21:* 216-228, 1999.

CRAWFORD JR and GARTHWAITE PH. Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia, 40*: 1196-1208, 2002.

CRAWFORD JR and GARTHWAITE PH. Statistical methods for single-case research: Comparing the slope of a patient's regression line with those of a control sample. *Cortex, 40*: 533-548, 2004.

CRAWFORD JR, GARTHWAITE PH and GRAY CD. Wanted: Fully operational definitions of dissociations in single-case studies. *Cortex, 39*: 357-370, 2003.

CRAWFORD JR and HOWELL DC. Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist, 12*: 482-486, 1998.

CRAWFORD JR, HOWELL DC and GARTHWAITE PH. Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired samples *t* test. *Journal of Clinical and Experimental Neuropsychology, 20*: 898-905, 1998.

D'AGOSTINO RB, BELANGER A and D'AGOSTINO RB. A suggestion for using powerful and informative tests of normality. *American Statistician, 44:* 316-321, 1990.

D'AGOSTINO RB and PEARSON ES. Tests of departures from normality. Empirical results for the distribution of *b2* and *b1*. *Biometrika, 60*: 613-622, 1973.

FOLSTEIN MF, FOLSTEIN SE and MCHUGH PR. 'Mini-mental state': a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research 12*: 189-198, 1975.

FUNG TD, CHERTKOW H, MURTHA S, WHATMOUGH C, PELOQUIN L, WHITEHEAD V and TEMPLEMAN FD. The spectrum of category effects in object and action knowledge in dementia of the Alzheimer's type. *Neuropsychology, 15*: 371-379, 2001.

GARRARD P, LAMBON-RALPH MA, WATSON PC, POWIS J, PATTERSON K, and HODGES JR. Longitudinal profiles of semantic impairment for living and nonliving concepts in dementia of Alzheimer's type. *Journal of Cognitive Neuroscience 13*: 892-909, 2001.

GARRARD P, PATTERSON K, WATSON PC and HODGES JR. Category specific semantic loss in dementia of Alzheimer's type. Functional–anatomical correlations from cross-sectional analyses. *Brain 121*: 633-646, 1998.

GONNERMAN LM, ANDERSON ES, DEVLIN JT, KEMPLER D and SEIDENBERG MS. Double dissociation of semantic categories in Alzheimer's disease. *Brain and Language 57*: 254-279, 1997.

LAIACONA M, BARBAROTTO R and CAPITANI E. Semantic category dissociations in naming: is there a gender effect in Alzheimer's disease? *Neuropsycholgia, 36*: 407-419, 1998.

LAWS KR. 'Illusions of Normality': a methodological review of category-specific naming. in press

LAWS KR. Sex differences in lexical size across categories. *Personality and Individual Differences, 36*: 23-32, 2003.

LAWS KR. Category-specific naming and modality-specific imagery. *Brain and Cognition, 48*: 418-420, 2002.

LAWS KR. Category-specific naming errors in normal subjects: the influence of evolution and experience. *Brain and Language, 75*: 123-133, 2000.

LAWS KR. Gender affects latencies for naming living and nonliving things. *Cortex, 35*: 729-733, 1999.

LAWS KR. A leopard never changes its spots. *Cognitive Neuropsychology, 15*: 467-479, 1998.

LAWS KR and GALE TM. Category-specific naming and the 'visual' characteristics of line drawn stimuli. *Cortex, 38*: 7-21, 2002a.

LAWS KR and GALE TM. Why are our similarities so different? A reply to Humphreys and Riddoch. *Cortex, 38*: 7-21, 2002b.

LAWS KR, LEESON VC and GALE TM. The effect of 'masking' on picture naming latencies. *Cortex, 38*: 137-147, 2002.

LAWS KR, LEESON VC and GALE TM. A domain-specific deficit for foodstuffs in patients with Alzheimer's disease. *Journal of the International Neuropsychological Society, 8*: 956-957, 2002.

LAWS KR, LEESON VC and GALE TM. Inflated and contradictory category naming deficits in Alzheimer's disease? *Brain and Cognition 53*: 416-418, 2003.

LAWS KR and NEVE C. A 'normal' category-specific advantage for naming living things. *Neuropsychologia 37*: 1263-1269, 1999.

MAURI A, DAUM I, SARTORI G, RIESCH G and BIRBAUMER N. Category-specific semantic impairment in Alzheimer's disease and temporal lobe dysfunction: a comparative study. *Journal of Clinical and Experimental Neuropsychology, 16*: 689-701, 1994.

MCKENNA P. *Category-Specific Names Test*. Psychology Press 1997.

MCKHANN G, DRACHMAN D, FOLSTEIN M, KATZMAN R, PRICE D and STADKLAN EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34*: 939-944, 1984.

MONTANES P, GOLDBLUM MC, and BOLLER F. The naming impairment of living and nonliving items in Alzheimer's disease. *Journal of International Neuropsychological Society, 1*: 39-48, 1995.

MYCROFT RH, MITCHELL DC and KAY J. An evaluation of statistical procedures for comparing an individual's performance with that of a group of controls. *Cognitive Neuropsychology, 19:* 291-299, 2002

PAYNE RW and JONES G. Statistics for the investigation of individual cases. *Journal of Clinical Psychology, 13:* 115-121, 1957

SHALLICE T. *From Neuropsychology to Mental Structure*. Cambridge UK: Cambridge University Press, 1988

SILVERI MC, DANIELE A, GIUSTOLISI L and GAINOTTI G. Dissociation between knowledge of living and nonliving things in dementia of the Alzheimer's type. *Neurology 41*: 545-546, 1991.

SNODGRASS JG and VANDERWART M. A standardised set of 260 pictures. *Journal of Experimental Psychology: Human Learning and Memory, 6*: 174-215, 1980.

TIPPETT LJ, GROSSMAN M and FARAH MJ. The semantic memory impairment of Alzheimer's disease: category specific? *Cortex, 32*: 143-153, 1996.

WALD A. An extension of Wilk's method for setting tolerance limits. *Annals of Mathematical Statistics*, 14, 45-55

ZANNINO GD, PERRI R, CARLESIMO GA, PASQUALETTI P and CALTRAGIRONE C. Category-specific impairment in patients with Alzheimer's disease as a function of severity: a cross-sectional investigation. *Neuropsychologia, 40:* 2268-2279, 2002.

Keith R Laws. Brain and Cognition Research Group, Division of Psychology. Nottingham Trent University, Burton Street. Nottingham, NG1 4BU, UK. e-mail: keith.laws@ntu.ac.uk

APPENDIX

Stimuli in Experiment 1

Banana	Airplane
Butterfly	Axe
Camel	Bicycle
Carrot	Bus
Cat	Chisel
Chicken	Drum
Corn	Flute
Cow	Glove
Frog	Guitar
Grapes	Helicopter
Mushroom	Kettle
Onion	Motorbike
Pear	Piano
Pineapple	Pliers
Spider	Rolling-pin
Squirrel	Screwdriver
Strawberry	Sledge
Duck	Stool
Ear	Tie
Elephant	Violin

Stimuli in Experiment 2a

Apple	Anchor
Banana	Axe
Camel	Barn
Carrot	Barrel
Cat	Basket
Celery	Bike
Cherry	Broom
Cow	Button
Dog	Chair
Duck	Chisel
Eagle	Cigar
Elephant	Couch
Fish	Crown
Fox	Drum
Frog	Glove
Goat	Hammer
Gorilla	Helicopter
Grapes	Kite
Kangaroo	Nail
Mouse	Needle
Onion	Peg
Ostrich	Pliers
Owl	Plug
Pear	Pram
Penguin	Rollerskate
Pig	Ruler
Pineapple	Sledge
Potato	Thimble
Rhino	Train
Squirrel	Umbrella
Tomato	Windmill
Zebra	Yacht

Stimuli in Experiment 2b

Mushrooms	Bat	Darts	Passport
Cucumber	Robin	Cracker	Calendar
Pineapple	Fox	Binoculars	Thermometer
Corn	Squirrel	Cue	Barrel
Rhubarb	Whale	Confetti	Cushion
Peach	Hedgehog	Whisk	Skittles
Cauliflower	Rhinoceros	Hand Grenade	Grate
Pepper	Eagle	Tambourine	Wreath
Celery	Eel	Triangle	Crate
Spring Onion	Badger	Mangle	Mould
Radish	Hippopotamus	Decanter	Weather Vane
Cress	Hare	Cymbals	Lantern
Garlic	Ostrich	Soda Syphon	Milk Churn
Marrow	Walrus	Plane	Barometer
Turnip	Gorilla	Plunger	Globe
Melon	Pheasant	Wash-Board	Snorkel
Broccoli	Flamingo	Crossbow	Bust
Beansprouts	Armadillo	Scuttle	Doily
Kiwi Fruit	Mole	Palette	Belisha Beacon
Courgettes	Platypus	Mallet	Cameo
Grapefruit	Porcupine	Ladle	Fez
Avocado	Hyena	Maracca	Casserole
Ginger	Vulture	Tankard	Boater
Aubergine	Otter	Kaleidoscope	Demijohn
Artichoke	Toucan	Parasol	Cauldron
Lychee	Puffin	Bugle	Seal
Mango	Beaver	Tureen	Jardinière
Chicory	Kiwi	Mortar and Pestle	Water Butt
Fennel	Lynx	Foil	Topi
Passion Fruit	Wildebeest	Besom	Tantalus