Determinants of Confrontation Naming Performance

Christopher Randolph
Loyola University Medical Center

Amy E. Lansing
Northwestern University Medical School

Robert J. Ivnik
Mayo Clinic

C. Munro Cullum
University of Texas Southwestern Medical Center

Bruce P. Hermann
University of Wisconsin Medical School

Confrontation naming tests are commonly employed in neuropsychological assessment. Surprisingly little work has been done, however, to determine how various demographic, linguistic, and disease status variables influence patterns of performance on these tests. The present study examined data on the Boston Naming Test (BNT) from a total of 1,131 subjects, including 719 normals, 325 patients with Alzheimer’s disease, and 87 patients with temporal lobe epilepsy. The effects of age, education, gender, and diagnostic group were examined with respect to overall scores, the influence of phonemic cuing, and performance on individual items. Profiles of scores on individual items were similar across diagnostic groups, suggesting that anomia is characterized by quantitative rather than qualitative changes in naming performance. Age and education systematically influenced scores. There was a significant effect of gender across diagnostic groups (males scoring higher than females), which appears to be due to performance on specific items. Phonemic cuing effects were similar across groups. The results are discussed with respect to the neuropsychological construct of confrontation naming and the clinical interpretation of performance on the BNT. © 1999 National Academy of Neuropsychology. Published by Elsevier Science Ltd

Address correspondence to Christopher Randolph, 111 North Wabash, Suite 1021, Chicago, IL 60602; E-mail: crandol@luc.edu
Confrontation naming tests, such as the Boston Naming Test (BNT; Kaplan, Goodglass, & Weintraub, 1983) are commonly employed in neuropsychological assessment. Surprisingly little work has been done, however, to determine how various demographic, linguistic, and disease status variables influence patterns of performance on these tests. The lack of such research has placed constraints on the clinical utility of these tests.

One variable which has been reported to systematically affect both the speed and accuracy of confrontation naming is the frequency with which the name of a given object appears in the lexicon. High-frequency items are typically named with greater speed and accuracy (Oldfield & Wingfield, 1965; Snodgrass & Vanderwart, 1980). The age at which a word is typically acquired in language development is another variable that has also been reported to affect naming latency. Words that are acquired early tend to be named more rapidly (Morrison, Ellis, & Quinlan, 1992). In fact, word frequency and age of acquisition are highly (negatively) correlated (Morrison et al., 1992). It is unclear which of these variables may be the primary determinant of naming performance, or whether they are in fact dissociable.

Age, education, and race may also be factors in naming performance, although the potential influence of these demographic variables has not been thoroughly investigated. An early study involving very high-functioning older adults (mean Full-Scale IQ = 122) reported only a mild decline on the BNT with aging (Van Gorp, Satz, Kiersch, & Henry, 1986). Another early study suggested relatively stable performance across age until the 70s, when performance was observed to decline (Albert, Heller, & Milberg, 1988). A more recent report involving a more heterogenous “cognitively intact” hospital population found considerably poorer performance in patients over 80 years of age (Ross, Lichtenberg, & Christensen, 1995). There is also some evidence that minority populations may perform more poorly on confrontation naming tests (Roberts & Hamsher, 1984; Ross et al., 1995), although sample sizes in these studies have been small and the effects of education not adequately addressed as a possible explanation for these differences. Gender differences, to our knowledge, have not been carefully explored, nor have anomic samples been directly compared to controls on standardized tests in terms of their pattern of failure across items.

The present study was undertaken to examine some of the potential determinants of performance on the BNT, which is probably the most commonly used clinical test of confrontation naming, comparing a large sample of normal controls to patients with Alzheimer’s disease and patients with temporal lobe epilepsy. The effects of age, education, gender, word frequency, and diagnostic classification were examined with respect patterns of performance on this test.

**METHOD**

Data on 719 normal controls, 325 patients with a diagnosis of Alzheimer’s disease (AD), and 87 patients with temporal lobe epilepsy (TLE) were gathered from four academic medical centers. AD patients were diagnosed as possible or probable AD according to NINCDS-ADRDA criteria (McKhann et al., 1984). TLE patients were referred for surgical evaluation because of medication-resistant epilepsy. All of the TLE patients were left-hemisphere dominant for speech, as determined by Wada testing. None of the TLE patients had any evidence of central nervous system lesion on MRI, apart from mesial temporal lobe atrophy. The normal controls consisted of both paid and unpaid volunteers for studies on neuropsychological functions in normal aging. Procedures for recruiting and further neuropsychological description of the majority of the normal sample
is detailed in previous articles (e.g., Ivnik et al., 1992; Randolph et al., 1994). Both clinical and normative samples were almost exclusively (>90% for each group) White.

Age, sex, education level (years), race, and diagnosis were coded for all subjects (see Table 1), as was their total score on the 60-item BNT, and their score on each of the 60 items. Items were scored as correct if they were correctly named spontaneously or after the stimulus (semantic) cue. Items correctly named after the phonemic cue were coded accordingly, but did not contribute to the total number correct. For the majority of the sample, the test was discontinued after 6 consecutive failures to name an item either spontaneously or after the stimulus cue (see below).

**RESULTS**

**Normative Data**

As this is the largest normative data set for the BNT that we are aware of in the literature, we felt that it would be appropriate to include a table of normative data (see Table 2). The data were broken down by age groups using the technique of overlapping midpoints, to conform to the Mayo Older Americans Normative Studies (e.g., Ivnik et al., 1992) for comparative purposes. After assuring that the BNT raw score distributions within each midpoint age interval were normal, we elected to present the data as means and standard deviations. The BNT norms in Table 2 are also broken down by gender,
given the significant differences between male and female performance (see below). Education effects are illustrated separately beneath Table 3, to avoid parcelling the data into unacceptably small cell sizes and excessively complicating the table.

**Completed Protocols**

The test was discontinued following 6 consecutive failures (correct naming after a phonemic cue was considered a failure) for all subjects except for 76 of the AD patients at one site, who were administered all 60 items regardless of number of consecutive failures. For these subjects, the average number of items they correctly named after failing 6 in a row was 2.0 (SD = 2.5; range 0–10). Although this cutoff rule seems therefore reasonably predictive of subsequent performance, given the nature of some of the intended comparisons in this study, it was felt that only those subjects who actually were administered all 60 items should be retained for the item profile and word frequency analyses (see below). This resulted in the retention of 717 NC subjects, 237 AD patients, and 50 TLE patients for those particular analyses.

**Phonemic Cuing**

The two patient groups did not differ in the number of additional items correctly identified following phonemic cuing (mean of 5.3 for AD patients, mean of 6.2 for TLE patients). NC subjects benefitted slightly, but significantly less than both patient groups, with a mean of 4.2 additional words correct. We felt that this modest attenuation of benefit of cuing in the normals was likely due to a ceiling effect.

**Age and Education Effects**

Age was inversely correlated with total scores in the NC group ($r = -0.45, p < 0.0001$). This correlation did not quite reach significance in the TLE group ($r = -0.2, p < 0.06$).
and was even less robust in the AD group ($r = -0.19, p < 0.001$), but was significant due to the large sample size. Education was positively correlated with performance in all three groups (NC: $r = 0.24$; AD: $r = 0.31$; TLE: $r = 0.44$; all $ps < 0.001$).

**Gender Effects**

A significant effect of gender was observed in all three groups on the total score, which remained after controlling for age and education through ANCOVA. Men outperformed the women in all three diagnostic groups (AD: $F = 27.5$; NC: $F = 42.9$; TLE: $F = 4.0$; all $ps < 0.05$). For the NC group, the effect of sex was explored for individual items, and it was determined that on 18 items men performed significantly better than women. The opposite was true for only 4 items (see Table 2). The preponderance of male-biased items appeared to produce the overall difference between the sexes.

**Item Profiles**

For each group, all subjects who had completed all 60 items entered this analysis (see above). The proportion of correct responses was calculated for each of the 60 items for each group (e.g., if 65% of the AD patients correctly named “dart,” that item would have a score of 0.65 for the AD group). These numbers, which represent the ease of naming each item for a particular group, were highly correlated among groups (Spearman rho): AD and NC: $r = 0.97$; AD and TLE: $r = 0.84$; NC and TLE: $r = 0.85$; all $ps < 0.0001$. This indicates that the profile of performance across items was highly similar across groups.

**Word Frequency Effect**

Word frequency values for each item were taken from Carroll, Davies, and Richman (1971), and correlated with the proportion of correct responses by group (as above). Word frequency values were significantly correlated with ease of naming in all groups (AD: $r = 0.46$; NC: $r = 0.41$; TLE: $r = 0.42$; all $ps < 0.002$). We were unable to obtain age-of-acquisition ratings on enough of the items to conduct a meaningful analysis.

**DISCUSSION**

The profiles of scores on individual items were highly similar across diagnostic groups, as measured by highly correlated naming rates for individual items across groups. This suggests that anomia is characterized by quantitative rather than qualitative changes in confrontation naming performance, at least for these two patient groups. This generalization seems likely to obtain across a variety of neurological disease states with diffuse involvement of temporal neocortex, however, given the marked differences between these two groups (e.g., age of disease onset, neuropathophysiology, age at testing).

For all groups, including controls, word frequency in the lexicon was significantly correlated with ease of naming. It is likely that age of acquisition, which is known to be highly correlated with word frequency, would be similarly related to naming performance. We were unable to find age-of-acquisition data on a sufficient number of items to conduct a meaningful analysis, however. Items with a very high frequency (and presumably low age of acquisition) were correctly named by virtually all subjects. This suggests a strategy for constructing future confrontation naming tests, by selecting primarily low-frequency stimuli. This can shorten test length by eliminating items that fail to con-
tribute to total score variance, as well as serve to eliminate the ceiling effects which are fairly common in tests of this type.

It is also noteworthy that the effects of phonemic cuing were similar across patient groups and normal controls, suggesting that this technique is not diagnostically useful, at least with these patient populations. It is unclear under what circumstances phonemic cuing might be of clinical utility. As it does not currently contribute to the total score on the BNT, these results suggest that it might be dispensed with under most circumstances. It has also recently been observed that some centers count correct responses following phonemic cuing as “correct” for the purpose of the discontinuation rule (Ferman, Ivnik, & Lucas, 1998). This variation in the application of the discontinuation rule can result in total score variation, which may be further reason for dispensing with this cuing format.

Age and education systematically influenced scores across groups. It is very important to note that our normal control group was a relatively homogeneous, almost entirely White, fairly well-educated group. It has been suggested that greater variance in performance among normals would be observed in a more culturally and racially heterogeneous sample (e.g., Roberts & Hamsher, 1984). This underscores the need for cautious interpretation of performance by any patient from a population that is inadequately represented in normative data on confrontation naming tests.

The variance associated with mean performance on the BNT increased as mean scores decreased; this was true both as a function of normal aging and with the presence of anomia associated with a disease state. The standard deviation associated with the normal mean for a midpoint age of 88 was approximately double that of the standard deviation associated with the normal mean for a midpoint age of 55. This is a fairly common finding in cognitive tests results across this age span (including the “old old”). In this case, it is unclear whether the increased variance simply represents the greater range of scores available as the mean moves away from the ceiling, or whether the increased variability should be interpreted as indicating that old age can be considered a disease state of sorts. The latter explanation might be supported by the hypothesis that, in some cases, “diminished reserve” leads to an accentuated effect of normal aging on cognitive abilities. A third explanation for this finding would be a cohort effect, with greater variability in the “old old” representing a greater developmental variability in verbal knowledge acquisition. Only longitudinal studies of neurocognitive functioning in the aged are capable of definitively determining which of these hypotheses is likely to account for the present findings, but in the interim, the results indicate the need for age-specific norms on tests of this nature.

One very surprising finding that has not, to our knowledge, been previously reported is the effect of gender on performance. This was a consistent and nontrivial effect size of about .5 for each diagnostic group, with males scoring higher than females. This is a reversal of the traditional finding of female superiority on verbal tasks. It appears to be due, however, to gender differences on specific items. Apparently, a number of items were selected for the BNT that are more salient (for some reason) for men, and therefore more easily named by men. It is interesting that some prior reports of Alzheimer’s patient performance on the BNT have found women performing worse than men, and been interpreted as reflecting greater severity of language impairment in female AD patients (e.g., Ripich, Petrill, Whitehouse, & Ziol, 1995). It is unclear from the present data whether that interpretation is tenable, as the magnitude of the difference (in terms of effect size) between men and women was similar for all groups.

As an aside, this finding also suggests that stimulus selection in test construction is worthy of closer scrutiny in order to avoid sources of systematic bias. For example, the reported female superiority in the CVLT normative reference sample has been inter-
Interpreted as evidence for a female advantage in verbal memory functions (Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology, 1996). This is an effect not observed in other standardized tests of verbal memory (e.g., The Wechsler Memory Scale-Revised; Wechsler, 1987). It seems possible from the present data, therefore, that this advantage was stimulus driven and therefore neither strictly language based or memory based. A different selection of stimulus items on the CVLT might conceivably result in superior performance by men.

Although the results of this study suggest strong commonalities underlying the pattern of performance on confrontation naming in different diagnostic groups, there are at least a couple of obvious exceptions to this rule. The first is a naming failure that is secondary to a perceptual deficit, as might be observed in a patient with a severe visuospatial perceptual impairment, or a visual object agnosia. Language cannot be reliably examined via visual confrontation naming in such patients, and their poor performance on such tests is not accurately described as anomic.

A more pertinent exception is observed in the rare patient with a “category-specific anomia.” These are cases with a demonstrated selective anomia for items within a specific semantic category (e.g., fruits and vegetables; Hart, Berndt, & Caramazza, 1985). What these unusual clinical presentations suggest is that the neural networks that constitute the stored representations of objects are semantically organized. Recent in vivo imaging research has supported this interpretation, suggesting that the brain regions activated during object identification are dependent upon the intrinsic properties of the object. In one study (Martin, Wiggs, Ungerleider, & Haxby, 1996), naming tools was found to activate both a portion of the left temporal lobe as well as a region in the left premotor area that is also activated by imagined hand movements. This was in addition to the activation of bilateral ventral temporal lobes and of Broca’s area that was observed in the naming of animals.

Future evolution of confrontation naming tasks may take advantage of this increasingly sophisticated understanding of the brain systems which mediate these functions, and tap distinct associative networks (e.g., action/nonaction naming). This would be of some particular interest in evaluating patients with focal disease processes. For the vast majority of patients who undergo neuropsychological evaluation, however, the underlying pathology is more diffuse. In the present study, both degenerative (AD) and developmental (epilepsy) diseases involving the temporal cortex appear to have had remarkably similar effects on patterns of confrontation naming performance as measured by the BNT.

Overall, the results suggest that anomia as measured by the BNT is best characterized as an exaggeration of normal naming failure, which is predicted in part by item frequency in the lexicon and is also dependent on demographic variables. These findings should inform clinical interpretation of performance on the BNT as well as attempts to construct, derive, or standardize new confrontation naming tests.

REFERENCES


