JSLHR

Article

Treatment of Category Generation and Retrieval in Aphasia: Effect of Typicality of Category Items

Swathi Kiran,[a] Chaleece Sandberg,[a] and Rajani Sebastian[b]

Purpose: Kiran and colleagues (Kiran, 2007, 2008; Kiran & Johnson, 2008; Kiran & Thompson, 2003) previously suggested that training atypical examples within a semantic category is a more efficient treatment approach to facilitating generalization within the category than training typical examples. In the present study, the authors extended previous work examining the notion of semantic complexity within goal-derived (ad hoc) categories in individuals with aphasia.
Methods: Six individuals with fluent aphasia (age range = 39–84 years) and varying degrees of naming deficits and semantic impairments were involved. Thirty typical and atypical items, each from 2 categories, were selected after an extensive stimulus norming task. Generative naming for the 2 categories was tested during baseline and treatment.
Results: As predicted, training atypical examples in the category resulted in generalization to untrained typical examples in 5 of 5 patient–treatment conditions. In contrast, training typical examples (which was examined in 3 conditions) produced mixed results. One patient showed generalization to untrained atypical examples, whereas 2 patients did not show generalization to untrained atypical examples.
Conclusion: Results of the present study supplement existing data on the effect of a semantically based treatment for lexical retrieval by manipulating the typicality of category examples.

Key Words: aphasia, rehabilitation, semantic, typicality

A central theme in categorization research has been the examination of graded structure in categories. *Graded structure* refers to the continuum of category representativeness, beginning with the most typical members of a category and continuing through its atypical members to those nonmembers least similar to category members. This idea of "graded membership" within categories was supported by several studies that showed differences in lexical processing between typical and atypical examples of a category. Typical examples generally receive preferential processing compared with atypical examples, and this phenomenon has been labeled the *typicality effect* (Hampton, 1979; Posner & Keele, 1968; Rosch, 1973, 1975). It turns out that common taxonomic categories such as *fruits* and *birds* all have graded structures. In fact, Vigliocco and colleagues argued that semantic distance (determined by the amount of feature overlap) is a stronger predictor of category organization and lexical access than category boundary definitions (Vigliocco, Vinson, Damian, & Levelt, 2002). Similar to common categories, goal-derived ad hoc categories also possess graded structures in which typicality can be determined for members of a particular category (Barsalou, 1983, 1985). These categories are instrumental to the achievement of goals—particularly, goals of daily living, such as *things at a garage sale*. However, ad hoc categories are more graded than common categories because they do not have rigidly defined features that constitute category membership. Instead, category members follow a loosely combined thread of common features. In addition, these categories are not as established in memory as common categories because people have had less experience with them as categorical concepts. That is, one does not specifically think about *things at a garage sale* as a discrete group of instances or a defined entity very often. Further, typical examples in goal-derived categories are those that are most suited to achieving the specific goal and are not necessarily the ones that share the most common properties. Despite this, typicality effects have been observed in ad hoc categories in healthy individuals as well as in patients with aphasia. In a series

[a]Boston University, Boston, MA
[b]University of Texas at Austin
Correspondence to Swathi Kiran: kirans@bu.edu

Editor: Robert Schlauch
Associate Editor: Lewis Shapiro
Received May 3, 2010
Revision received October 7, 2010
Accepted November 21, 2010
DOI: 10.1044/1092-4388(2010/10-0117)

of studies, Hough (1993) investigated patients' awareness and knowledge of goal-derived and common category structure. Hough found that fluent and nonfluent patients with aphasia exhibited typicality patterns similar to those of controls on a category generation task for goal-derived categories; however, these patients were more anchored to the central portion of a category's referential field for common categories. On a category verification task, Hough found that fluent and nonfluent patients required significantly more time to identify category examples than did the non-brain-damaged individuals; however, no differences between these two groups were observed in their overall typicality pattern. Therefore, these results suggest that patients with aphasia exhibit relatively unimpaired representations of goal-driven categories and are sensitive to the graded effects within such categories. In addition, in previous work, we also found that patients with aphasia responded faster to typical examples than to atypical examples in ad hoc categories. Healthy young and older participants and patients with aphasia participated in an online category verification task in which *primes* were ad hoc category labels and *targets* were typical members, atypical members, nonmembers, or nonwords. All three groups were significantly faster and more accurate on typical examples than on atypical examples; however, patients with semantic impairment differed from their nonsemantically impaired counterparts and from healthy controls by showing abnormal typicality effects for ad hoc categories (Sandberg, Sebastian, & Kiran, 2011). Therefore, these studies seem to be inconclusive about the representation of typicality in ad hoc categories across different types of patients with aphasia. Further, the extent to which the graded nature of category representation can be exploited in treatment for lexical retrieval in patients with aphasia has not been examined. This is an important empirical question because ad hoc categories have a distinct advantage in that the number of items that can be retrieved for a category label is potentially endless and, therefore, may be more conducive to examining lexical retrieval than a traditional picture-naming task.

Thus, the present study was aimed at manipulating typicality (or gradedness) as a treatment variable to facilitate lexical retrieval in individuals with aphasia. In our previous work, we suggested that training the more complex atypical examples in semantic categories results in generalization to the less complex typical examples because although atypical items are less representative of their category than are typical items, features common to both typical and atypical items are trained. These findings have borne out in three studies that examined generalization from atypical to typical examples and vice versa using animate categories (*birds, vegetables*; Kiran & Thompson, 2003), inanimate categories (*clothing, furniture*; Kiran, 2008), and well-defined categories (*shapes*; Kiran & Johnson, 2008) and have formed the basis for the complexity hypothesis (Thompson, 2007) within the semantic domain (Kiran, 2007). This idea of complexity extends into domains other than categories of concrete objects. Similar findings were observed during treatment of category generation of abstract or concrete words by using abstractness as a marker of complexity. Three of four patients with aphasia showed improvement on trained abstract words (e.g., *prayer* within the context of *church*) and generalization to untrained concrete words (e.g., *candle* within the context of *church*). Two of the four patients showed improvement on trained concrete words, but no generalization was observed to untrained abstract words (Kiran, Sandberg, & Abbott, 2009).

In the present study, we hypothesized that, similar to semantic categories such as *birds* and *furniture,* ad hoc/goal-derived categories such as *things at a garage sale* are represented in terms of typicality, with typical items (e.g., *cups, clocks*) in the center of the semantic space and atypical items (e.g., *candles, pens*) at the periphery of this space. Therefore, we hypothesized that training atypical items in a category would result in generalization to untrained target typical items in the same category. However, we also hypothesized that training typical items in a category would result in the retrieval of trained typical items but not generalization to target atypical items. An important assumption here was that words in ad hoc categories are represented as semantic features (McRae, de Sa, & Seidenberg, 1997; Pexman, Holyk, & Monfils, 2003) and that training semantic features for a specific word should improve access to the phonological form of the word and its semantically related neighbors (Kiran & Bassetto, 2008). With that in mind, the theoretical premise proposed in this study was that training items at the periphery would strengthen a more distributed set of featural representations of items that help fulfill the goal of the category, whereas training featural representations of items at the center of the category would reinforce only the core features that fulfill the goal but not the featural variations. For instance, semantic features relevant to atypical examples such as *used as protection from weather, multipurpose,* and *used for personal hygiene* fulfill the goals of a goal-derived category such as *things to take camping* and also represent the featural variation of the category. We expected that training such semantic features for relevant atypical examples would affect a broader range of examples within the category than would training features such as *used for setting up camp,* which only fulfill the core goal of *things to take camping*.

As another goal of the study, we examined whether training patients to generate typical or atypical examples for a category would likely influence their ability to generate other semantically related examples within the category. Therefore, two questions were posed. First, if strengthening semantic features relevant to atypical examples results in strengthening access to words that are represented by a greater variety of features, could it be speculated that the training of access to atypical examples within ad hoc categories results in improved access to a larger number of examples within the category than does the training of typical examples? To address this question, we evaluated the total number of responses produced as a function of treatment (typical vs. atypical treatment). Second, will patients develop and implement different strategies during improved category generation as a function of treatment? To address this question, we categorized responses that were generated into subgroups (clusters) and examined whether or not patients showed trends in generating similar semantic clusters.

From a clinical standpoint, examining ad hoc (or goal-derived) categories also permitted us to extend the typicality treatment protocol beyond confrontation picture naming to a category generation task. We successfully implemented the category generation task as a dependent variable in a previously described study examining complexity (Kiran et al., 2009). Because ad hoc categories differ from common language categories in that they lack distinct category boundaries, a potentially vast list of examples can be generated for each category examined. Consequently, a category generation task is more conducive to examining lexical retrieval within such categories than is picture naming. Interestingly, Hough (2007) found that both middle-aged and older healthy participants generated similar numbers of responses for goal-derived categories (but not for common language categories) as compared to young participants, reflecting the fact that the ability to generate associative connections for goal-derived categories does not decline with increasing age. Therefore, using a category generation task for ad hoc (goal-derived) categories provides a practical and naturalistic opportunity to assess lexical retrieval in patients with aphasia. To summarize, in the present study we used a category generation task to facilitate lexical retrieval in patients with aphasia and examined whether or not manipulating the typicality of items within specific goal-derived categories resulted in selective acquisition and generalization patterns for trained and untrained items.

Method

Participants

Six monolingual, English-speaking individuals with aphasia were recruited from local hospitals within the Austin, Texas, area to participate in the study. Several initial selection criteria were met, including (a) a single left-hemisphere stroke in the distribution of the middle cerebral artery confirmed by a CT/MRI scan, (b) onset of stroke at least 6 months before participation in the study, (c) premorbid right-handedness as determined by a self-rating questionnaire, and (d) at least a high school diploma (see Table 1). All participants also passed an audiometric hearing screening at 40 db HL bilaterally at 500, 1000, and 2000 Hz and showed normal or corrected-to-normal vision as measured by the Snellen chart. All participants had received varying amounts of traditional language treatment during the initial months after their stroke but were not involved in any concurrent therapy during the study. All participants provided written consent approved by the University of Texas Institutional Review Board.

The diagnosis of aphasia was determined by administration of the Western Aphasia Battery (WAB; Kertesz, 1982). All participants were fluent, anomic, or conduction aphasic (see Table 1 for details). Several other inclusionary criteria for participation in the study were used. First, performance on the Boston Naming Test (BNT; Goodglass, Kaplan, & Weintraub, 1983) was required to be below 65% (40 of 60 points) accuracy to ensure that participants showed lexical retrieval impairments. However, all participants could name at least some of the pictures on the BNT, indicating that they did not demonstrate a severe naming impairment. Next, all participants demonstrated impaired category generation on the WAB category fluency task, which was deemed to be similar to the

Table 1. Demographic and stroke-related data for the 5 study participants.

Characteristic	P1	P2	P3	P4	P5	P6
Age (yrs)	76	39	76	69	84	64
Gender	F	F	M	M	F	M
Handedness	Right	Right	Right	Right	Right	Right
Occupation	Teacher	Software engineer	Retired clerk	Business	Clerk	Attorney
Etiology	Left CVA	Left TP hemorrhage	Left CVA	Left TP CVA	Left TP CVA	Left CVA
MPO	30	6	108	10	9	96
Aphasia quotient	79	82	84.3	72.1	70.9	84.8
Aphasia diagnosis	Anomic	Anomic	Anomic	Conduction	Conduction	Anomic

Note. yrs = years; P = participant; F = female; M = male; CVA = cerebrovascular accident; TP = temporoparietal; MPO = months postonset.

treatment-dependent variable (see Table 2). Except for Participant 4 (P4), all participants demonstrated mild semantic impairments on four semantic processing subtests that were administered from the Psycholinguistic Assessment of Language Processing in Aphasia test (PALPA; Kay, Lesser, & Coltheart, 1992) and the Pyramids and Palm Trees test (PAPT; Howard & Patterson, 1992). With the exception of P4, all patients showed mild impairments in written naming but fairly accurate reading and repetition skills, validating the hypothesis that patients demonstrated lexical retrieval impairments and not impairments in the phonological output lexicon. P4 presented with a combination of semantic and phonological impairments.

With the exception of P1, all participants were also administered the Cognitive Linguistic Quick Test (Helm-Estabrooks, 2001). All participants who were tested performed either within normal limits or with mild impairments on all components of the test. Exceptions to this included P5 showing moderate impairment on the language component and P4 having significant difficulty with the memory, language, and attention components (see Table 3).

To assist in the development of norms for the stimuli used in the study, 20 young (age range = 21–40 years) and 20 older (age range = 41–75 years) individuals were recruited from the University of Texas at Austin. All participants had normal or corrected-to-normal vision, normal hearing, and at least a high school diploma. Exclusionary criteria included history of neurological disorders, psychological illnesses, alcoholism, learning disability, seizures, and attention-deficit disorders.

Stimuli

Two ad hoc categories and their examples were developed for use in the experiment. In the stimulus development and norming phase, 20 healthy young and older individuals generated as many items as possible for five ad hoc categories (*things to take camping, things at a grocery store, things at a garage sale, things that fly,* and *things that smell*). The items generated were entered into a database of words using the Linguistic Inquiry and Word Count (LIWC) software (http://www.liwc.net/index.php) that is described in greater detail below. These categories were used in previous ad hoc category studies (Barsalou, 1983, 1985; Hough, 1993). A separate group of 20 healthy young and older individuals rated the previously generated items for each of the five ad hoc categories on a 7-point scale. A rating of 1 corresponded to the item being a *very good example or fit of the category*; a

Table 2. Performance on the WAB, BNT, PALPA, and PAPT.

Test	P1		P2		P3		P4		P5		P6	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
WAB												
Spontaneous Speech	19	20	9	9	16	12	16	16	16	19	15	16
Auditory Comprehension	8.3	8.35	10	10	9.15	7.6	8.05	7.05	7.25	6.6	10	9.8
Repetition	3.9	6.6	7.2	8	8.9	8.4	5.7	5.6	5.6	8.2	8.2	9.5
Naming	8.3	7.9	5.8	7.5	8.1	8.2	6.3	5.8	6.6	5.7	9.2	9.5
Category Fluency	12	9	4	10	6	7	7	5	12	6	16	16
Aphasia Quotient	79	85.7	82	89	84.3	85.6	72.1	68.9	70.9	73.1	84.8	89.6
BNT (%)	43.3	58.3	21.7	23.3	68.3	66.7	18.3	15.0	26.7	25.0	56.7	53.3
PALPA												
Auditory Lexical Decision (%)	88.1	90.6	98.8	94.4	93.8	85.0	88.8	92.5	73.8	80.6	96.9	96.3
Letter Length Reading (%)	100.0	100.0	100.0	100.0	100.0	100.0	50.0	54.2	87.5	100.0	70.8	75.0
Visual Lexical Decision Task (%)	93.3	91.7	93.3	94.2	90.0	91.7	73.3	77.5	90.0	78.3	98.3	99.2
Spoken Word–Picture Matching (%)	100.0	100.0	100.0	97.5	100.0	100.0	95.0	92.5	90.0	95.0	100.0	100.0
Written Word–Picture Matching (%)	100.0	100.0	100.0	100.0	100.0	97.5	97.5	97.5	95.0	92.5	100.0	100.0
Auditory Synonym Judgments (%)	88.3	90.0	85.0	86.7	95.0	86.7	60.0	61.7	65.0	65.0	90.0	90.0
Written Synonym Judgments (%)	100.0	96.7	86.7	91.7	95.0	85.0	56.7	58.3	73.3	71.7	91.7	95.0
Spoken Picture Naming (%)	100.0	100.0	92.5	95.0	95.0	92.5	62.5	60.0	62.5	70.0	90.0	95.0
Writing Picture Names (%)	80.0	85.0	85.0	92.5	92.5	100.0	7.5	132.5	62.5	92.5	92.5	95.0
Reading Picture Names (%)	97.5	100.0	97.5	100.0	100.0	95.0	37.5	52.5	92.5	0.0	85.0	95.0
Spelling Picture Names (%)	92.5	92.5	95.0	95.0	85.0	85.0	10.0	0	82.5	60.0	95.0	97.5
PAPT–3 Pictures (%)	96	96	92	90	96	92	92	94	86	DNT	96	96

Note. WAB = Western Aphasia Battery; BNT = Boston Naming Test; PALPA = Psycholinguistic Assessment of Language Processing in Aphasia; PAPT = Pyramids and Palm Trees test.

Table 3. Performance on the Cognitive Linguistic Quick Test (CLQT).

CLQT	P1		P2		P3		P4		P5		P6	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Attention		WNL	WNL	WNL	WNL	WNL	Mod	Mod	WNL		WNL	WNL
Memory		WNL	WNL	Mild	WNL	WNL	Severe	Severe	Mild		WNL	WNL
Executive Functions		WNL	WNL	WNL	Mild	WNL	Mod	Mild	WNL		WNL	WNL
Language		Mild	Mild	Mild	Mild	WNL	Severe	Severe	Mod		WNL	WNL
Visuospatial Skills		WNL	WNL	WNL	WNL	WNL	Mild	Mild	WNL		WNL	WNL

Note. WNL = within normal limits; Mod = moderate. Blank cells indicate "did not test."

rating of 7 indicated that the item was considered a *very poor example*; a rating of 4 indicated a *moderate fit* (Rosch, 1975). After the participants completed the task, the average rating score, *SD*s, and *z*-scores for each item were calculated across the 20 participants. Items that were eliminated included (a) those whose average typicality rating occurred with an *SD* of 2 or more; (b) those that consisted of two or more synonyms in the list, of which one was a superordinate label; (c) those that were both atypical and unfamiliar; and/or (d) those that were verbs. After specific items in each category were deleted, some categories were left with too few items to separate into typical and atypical groups. As a result, the following categories were eliminated: *things that fly, things at a grocery store,* and *things that smell*. The two remaining categories (*things at a garage sale* and *things to take camping*) were selected for treatment.

Stimuli for each experimental category were selected on the basis of the *z*-scores of the average typicality ratings for each rated item. The 15 items with the lowest *z*-scores were considered typical examples (e.g., *garage sale* = –0.50, *camping* = –0.55), and the 15 items with the highest *z*-scores were considered atypical examples (e.g., *garage sale* = 0.62; *camping* = 0.68). We made every effort to ensure that there were no differences in the written word frequency between the typical and atypical examples (e.g., *garage sale* M_{typical} = 21.84; *garage sale* M_{atypical} = 19.28, $t = 0.37$, $p = .71$; *camping* M_{typical} = 17.36; *camping* M_{atypical} = 12.11, $t = 1.06$, $p = .30$) and familiarity (*garage sale* M_{typical} = 527, *garage sale* M_{atypical} = 554, $t = -0.57$, $p = .57$; *camping* M_{typical} = 540; *camping* M_{atypical} = 536, $t = 0.14$, $p = .88$; Coltheart, 1981). In some cases, we kept two-word phrases because they were an integral part of the representation of the category (e.g., *sleeping bag* for *camping*). Individual typed cards were printed for all words.

Development of semantic features for treatment. For each of the two categories, 10 healthy participants listed as many semantic features as possible. Instructions to participants were as follows:

> In this experiment, I will give you a written list of items and ask you to write down as many attributes as you can that you think are applicable for the given examples. Please keep in mind that there is no right or wrong response. Please provide at least 15–20 attributes that are relevant for all or some of the examples provided.

As expected, the *garage sale* category elicited many more features ($N = 46$) than did *camping* ($N = 22$). Then, the number of items that each feature applied to was tabulated. Certain features were applicable to all examples and fulfilled the goal of the category (e.g., *unwanted/unneeded* for *garage sale*) and were hence labeled "core" features. Others were applicable to typical examples in the category but were not considered integral to fulfilling the goal (e.g., *buy at a sporting goods store* for *camping*). Still others were applicable only for atypical examples in the category (e.g., *things needed for cooking* for *camping*). Similarly, the number of features that each example did or did not possess was also tabulated, ensuring that there was a relatively even distribution of features across the examples generated for the category. See Appendix A for a representative list of semantic attributes used for each of the two treatment categories. For each category, we developed 15 distracter features that did not apply to any of the examples in the category. Consequently, there were a total of 58 features for the *garage sale* category and a total of 31 features for the *camping* category. It was reasoned that the differential number of features selected for each category would not affect treatment outcomes; however, administration of a multiple-baseline, across-categories treatment design allowed for a systematic examination of this issue.

Design

This study used a single-subject, multiple-baseline, across-categories experimental treatment design with the order of category and typicality counterbalanced across the six participants (see Table 4). The criterion for switching treatment from one category to the next was set at either 80% accuracy in generating target items in the trained category on two consecutive treatment probes or the completion of 20 training sessions.

Table 4. Number of baselines and counterbalanced order of category and typicality exposed in treatment, with summary of acquisition and generalization effect sizes (ESs) for trained items and untrained items.

Participant	# of baselines	Category trained	Trained item acquisition	ES target atypical	ES other atypical	Untrained item generalization	ES target typical	ES other typical
1	3	1. Things at a garage sale	Atypical	20.5	17.6	Typical	9.8	21.4
2	4	2. Things to take camping	Atypical	7.6	0.6	Typical	3.6	0.5
3	3	1. Things at a garage sale	Atypical	11	4.6	Typical	10.4	4
4	14	1. Things at a garage sale	Atypical	14.3	1.7	Typical	2.4	2.6
6	3	1. Things at a garage sale	Atypical	10	8.1	Typical	2.6	1.7
2	6	1. Things at a garage sale	Typical	1.9	−1	Atypical	3	0.1
4	5	2. Things to take camping	Typical	14.7	0[a]	Atypical	−0.05	0[a]
5	3	2. Things to take camping	Typical	6.06	1.2	Atypical	1.2	0.9

Note. Acquisition benchmark was set at an ES of 6.5; generalization benchmark was ES = 2.0.

[a]In these cases, ESs could not be calculated due to the *SD* in the baseline equaling zero.

Baseline naming procedures. Generative naming for the two categories was tested during baseline. The number of baselines was varied across participants to evaluate the stability of the dependent variable. Participants were instructed to name as many words associated with each category/location as they could without a time limit. The number of target typical words (e.g., *tent, flashlight* for *camping*) and the number of target atypical words (e.g., *pen, pillow* for *garage sale*) were tabulated. Target typical and atypical words were category examples that were normed according to the procedures described above. They were marked as *generated* (1) or *not generated* (0) and were considered correct if they were clear and intelligible productions of the target words, semantically similar variations of the target word, or a very close synonym of the target word. We also kept track of (a) untrained typical words (e.g., *bread maker, skillet* for *garage sale*) and (b) untrained atypical items (*necklace* for *garage sale*), which were category examples for which we had typicality norms and that were spontaneously generated by each participant. These responses were considered to be correct untrained words only if they were intelligible productions of words that were appropriate for the category and that were identical or semantic variations of items for which we had typicality *z* ratings from our normed data set (e.g., *DVDs* counted as a correct alternative response for *movies*). Note that we could not classify all responses produced by each participant as untrained typical/atypical examples because our norms on typicality for items in the category are for approximately 50 items, whereas the participants generated many more items during the sessions. In addition, participants produced several items that were similar to the norms collected but did not carry the same meaning in the context of the category (e.g., *shoes/hiking boots*). Consequently, the data that were reported as untrained typical or atypical are somewhat subjective and should be interpreted in the context of this limitation. To circumvent this issue, we also tabulated all spontaneous generations that did not belong to the target typical/atypical set but were considered acceptable members of a category.

Qualitative analysis of responses. We also conducted a qualitative analysis of the responses generated. First, all responses generated by patients and healthy controls during the norming tasks were entered into a database using the LIWC software (http://www.liwc.net/index.php). We used LIWC to count the total number of items produced by all healthy controls and all patients for each category and assembled them into a "category dictionary." In this way, two dictionaries were created: (a) *garage sale,* with 596 unique words, and (b) *camping,* with 469 unique words. For each category, dictionary responses were categorized into subcategories (*camping: supplies/tools, clothing/personal care, food/cooking, games/entertainment, transportation, wildlife/animals,* and *other*; *garage sale*: *clothing, kitchen, electronics, fruits/vegetables, furniture, entertainment, prepared food/drinks, home/garden,* and *miscellaneous*). Then, we used LIWC to count the number of times that a word in a particular subcategory was produced. Responses that did not fall under these categories were classified as production errors.

Treatment. Each treatment session was carried out in four steps: (1) category generation, (2) category sorting, (3) feature generation/selection, and (4) yes/no feature questions (see Appendix B). Patients were seen two times per week for 2 hr each session.

Treatment probes. Throughout treatment, the same generative naming probes used in the baseline condition were presented every second treatment session to assess retrieval of the trained and untrained items. Generalized retrieval of untrained items was considered to have occurred when levels of performance changed by at least 40% over baseline levels.

Reliability. All baseline sessions and treatment sessions were recorded on videotape. Reliability on the dependent variable for participants was calculated for 75% of the probe sessions, resulting in 100% agreement. Reliability on the independent variable (i.e., presentation of the treatment protocol) was calculated for 50% of treatment sessions, resulting in 100% agreement.

Data analysis. To calculate effect sizes (ESs), we subtracted the average baseline probe scores from the average post-treatment scores and divided the result by the *SD* of the baseline scores (Beeson & Robey, 2006). In cases where treatment was provided for a second category, all pretreatment sessions were entered into the baseline calculation (e.g., for P4 *garage sale* treatment, 15 data points were entered into the baseline calculation). For P1 and P3, post-treatment probe scores could not be obtained as a result of patient evaluation scheduling issues. Consequently, for these two patients, we used the average of the final two treatment probe scores. Beeson and Robey (2006) recently updated the benchmarks for direct treatment of naming deficits and generalization of treatment (for acquisition of trained items, 6.5 = small ES, 8.0 = medium ES, and 9.5 = large ES; for generalization of treatment, 2.0 = small ES, 5.0 = medium ES, and 8.0 = large ES). In order to consider the treatment effective, we set a benchmark of ES = 6.5 for the trained items and an improvement to 80% accuracy for two consecutive sessions. Likewise, for generalization to be considered positive, we set a benchmark of ES = 2.0 and an improvement of 40% accuracy over baseline levels for the untrained items (see Table 4).

Results

Category Generation Treatment

P1. P1 received treatment for atypical examples of *garage sale* for 7 weeks. These items improved to criterion (see Figure 1 and Table 4), and generalization occurred for untrained typical words. Treatment was not provided for the second category as a result of personal health issues.

P2. P2 received 6 weeks of treatment for atypical examples of *camping*. Retrieval of the trained atypical items improved, and generalization also occurred for the untrained typical items of the category (see Figure 2 and Table 4). Treatment was then shifted to typical examples of *garage sale,* which improved to criterion but did not meet our a priori ES criterion. Interestingly, some generalization was observed to untrained atypical examples.

P3. P3 received 15 weeks of treatment for atypical examples of *garage sale*. Treatment was extended from 10 to 15 weeks for this patient because he showed trends of improvement on the trained items but an overall variable performance. This patient did eventually improve to 87% accuracy on the trained atypical items (see Figure 3 and Table 4). Generalization to the untrained typical items was modest in accuracy but yielded a large ES (see Table 4). This patient developed a health complication toward the end of treatment and chose not to continue in treatment.

P4. P4 was trained on typical examples of *camping* for 10 weeks. Retrieval of trained items improved, but no generalization was observed for the untrained atypical examples. Treatment was then shifted to atypical examples of *garage sale*. Again, improvement was noted, but the number of items retrieved did not reach criterion (see Figure 4 and Table 4). Notably, generalization to the untrained typical examples was observed, and importantly, P4 could retrieve more typical examples than trained atypical examples.

P5. P5 was trained on typical examples of *camping* for 10 weeks. Retrieval of trained items improved but did not reach criterion. Little to no generalization was

Figure 1. Percent of target responses produced for atypical (trained) and typical (untrained) items for the *garage sale* category. Treatment was not provided for the second category. P1 = Participant 1.

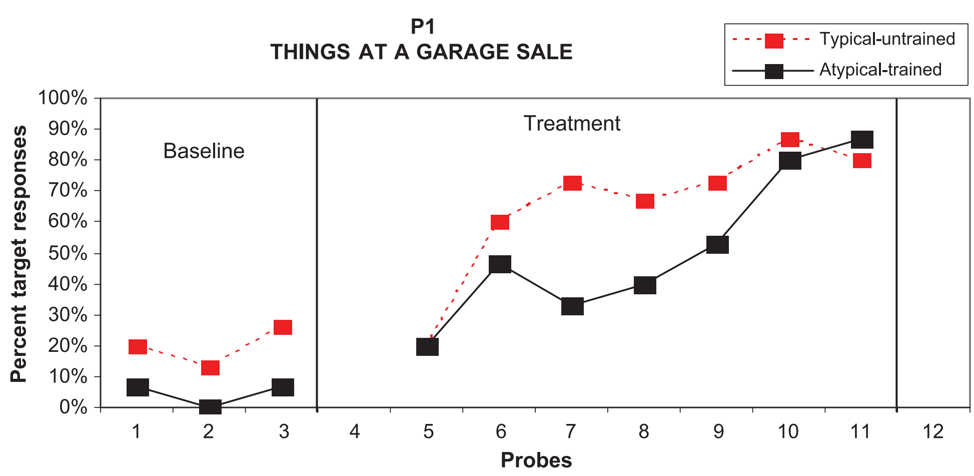


Figure 2. A: Percent of target responses produced for atypical (trained) and typical (untrained) items for the *camping* category for participant 2 (P2). B: Percent of target responses produced for typical (trained) and atypical (untrained) when treatment was provided for the *garage sale* category for P2.

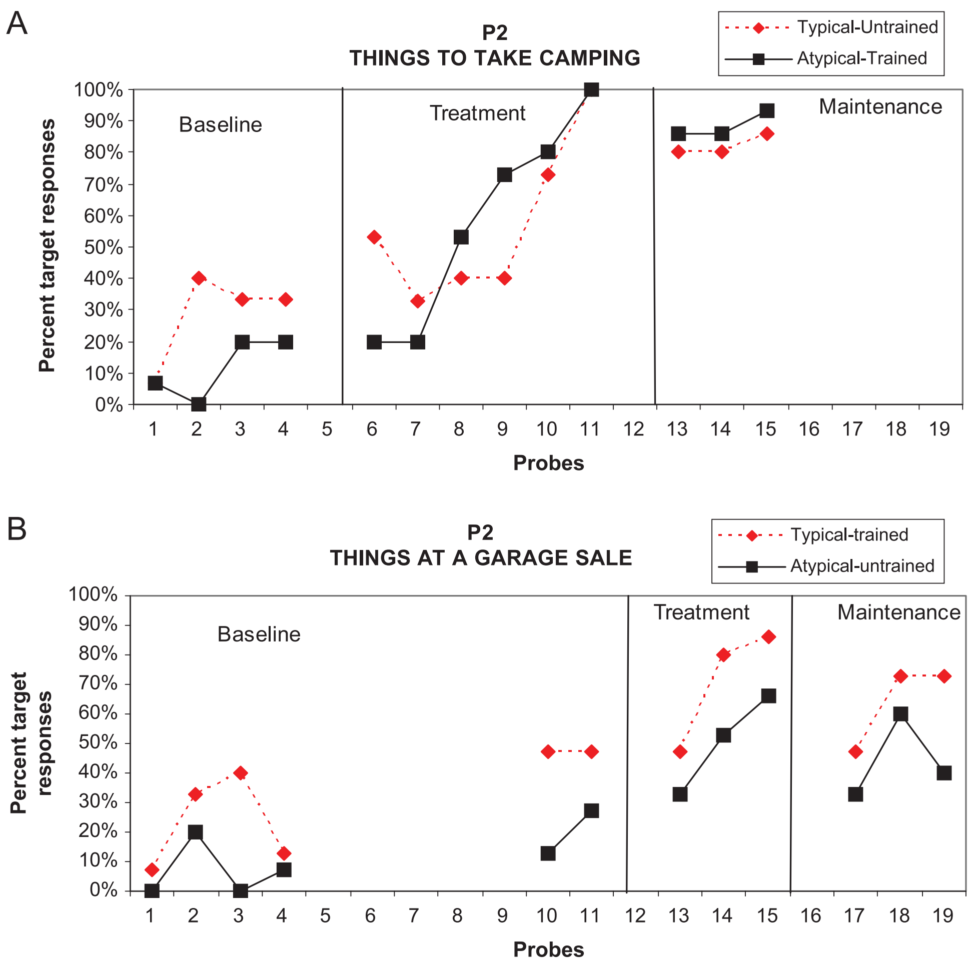


Figure 3. Percent of target responses produced for atypical (trained) and typical (untrained) items for the *garage sale* category for participant 3 (P3). Treatment was not provided for the second category.

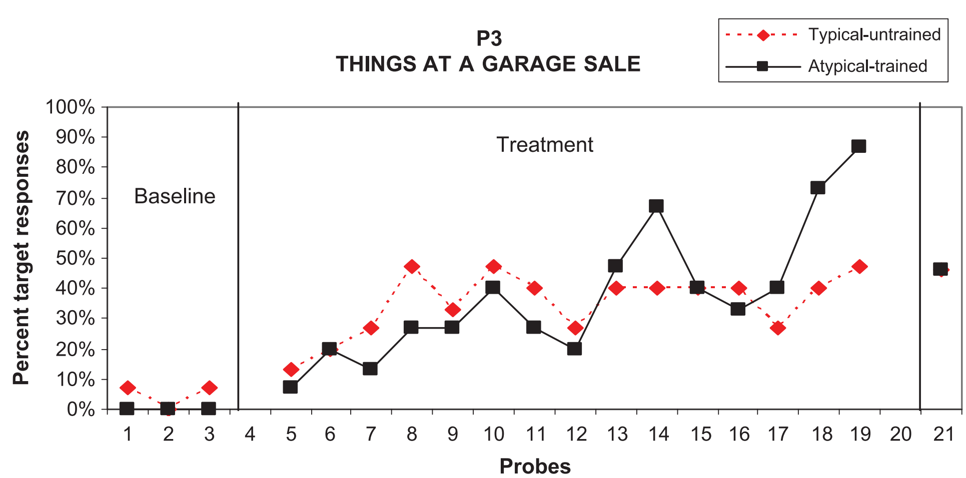
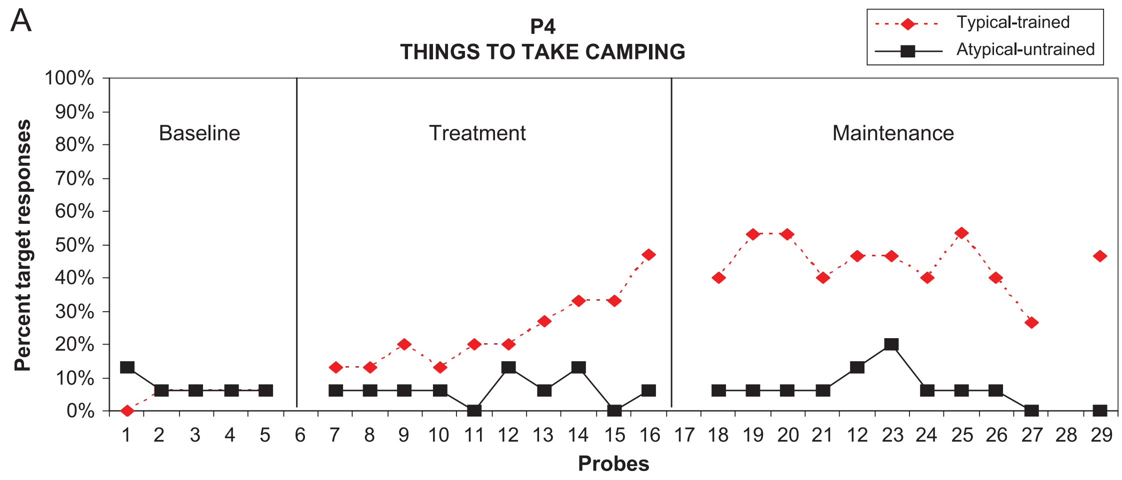


Figure 4. A: Percent of target responses produced for typical (trained) and atypical (untrained) items when treatment was provided for the *camping* category for participant 4 (P4). B: Percent of target responses produced for atypical (trained) and typical (untrained) when treatment was provided for the *garage sale* category for P4.



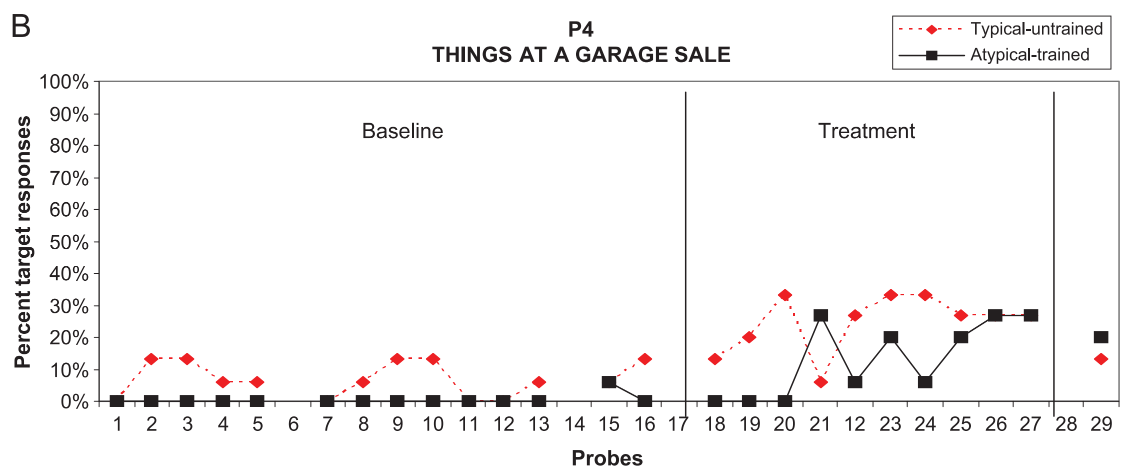
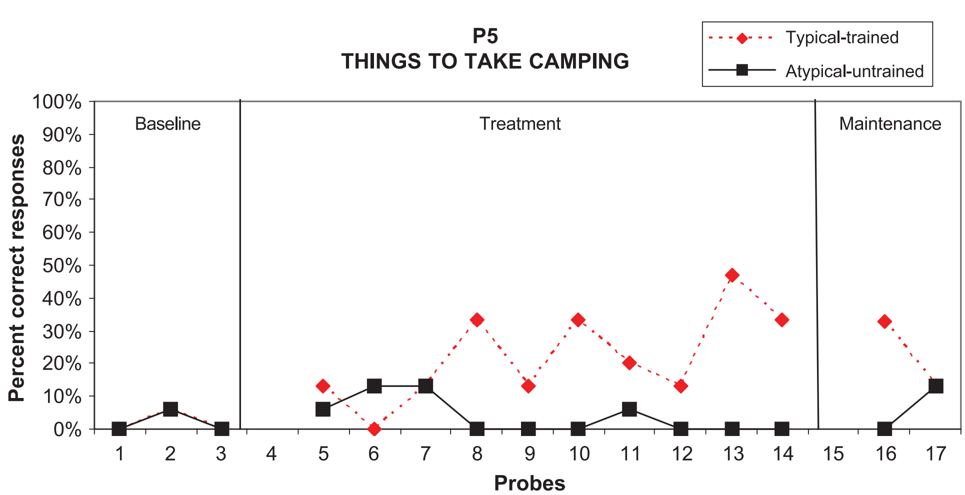


Figure 5. Percent of target responses produced for atypical (trained) and typical (untrained) items for the *garage sale* category for participant 5 (P5). Treatment was not provided for the second category.



observed for the untrained atypical examples (see Figure 5 and Table 4). P5 was not trained on the second category per the participant's wish to terminate treatment.

P6. P6 was trained on atypical examples of *garage sale* for 10 weeks. Retrieval of trained atypical items reached criterion. Generalization to untrained typical examples was also observed, with retrieval of typical examples reaching levels that were consistently higher than those for trained atypical examples during the course of treatment (see Figure 6 and Table 4). P6 also expressed an interest in terminating treatment as a result of scheduling conflicts.

Quantitative and Qualitative Analysis of Other Responses Generated for Each Category

The responses generated by each participant were further examined to see whether therapy influenced the nature of retrieval. For this analysis, target typical and atypical words were eliminated from the data set because these responses were directly related to the therapy and are already illustrated in Figures 1–6. Next, production errors, phonological errors, and/or neologisms were eliminated from the data set and were not analyzed further. In order to understand whether or not the trained stimuli (either typical or atypical) had any effect on the number of responses produced as a consequence of treatment, we conducted an analysis of the overall number of responses in the category. Responses generated by each participant during the first two baseline sessions and the final two treatment/post-treatment sessions were tabulated. For this analysis, we included only the total numbers of items in each category (and not the categorized numbers). A repeated measures analysis of variance using the average number of responses produced during baseline and end-of-treatment probes as the dependent variable and participant and typicality of treatment stimuli as the independent variables showed significance, Wilks' $\lambda = .02814$, $F(14, 14) = 4.9610$, $p = .00249$. Figure 7 shows that, in general, participants who were trained on atypical examples in a category generated more items for that category at the end of treatment than did participants who were trained on typical examples of a category.

Next, we regrouped all responses (including target trained, target untrained, other responses, and production errors) to see whether there were changes as a function of treatment. These examples belonged to one of several subcategories for the category (e.g., for the *camping* category, other responses could be examples of *supplies/tools, clothing/personal care, food/cooking, games/entertainment, transportation, wildlife/animals,* and *miscellaneous*), with the exception of target typical and target atypical examples. For each patient, the proportion of responses within each subtype was computed for each session. Then, for each subtype of response, we subtracted the average of the first two (baseline) sessions from the average of the final two (treatment/post-treatment) sessions to obtain a difference score. We decided to include data from two sessions in order to get a measure of consistency across error types. Negative values reflect a decrease in the proportion of response subtypes as a function of treatment, whereas positive values reflect an increase in the response subtypes subsequent to treatment. We conducted a hierarchical cluster analysis to see whether patients showed similar trends in their production of different response subtypes as a function of treatment. The difference score described above was entered into a hierarchical

Figure 6. Percent of target responses produced for atypical (trained) and typical (untrained) items for the *garage sale* category for participant 6 (P6). Treatment was not provided for the second category.

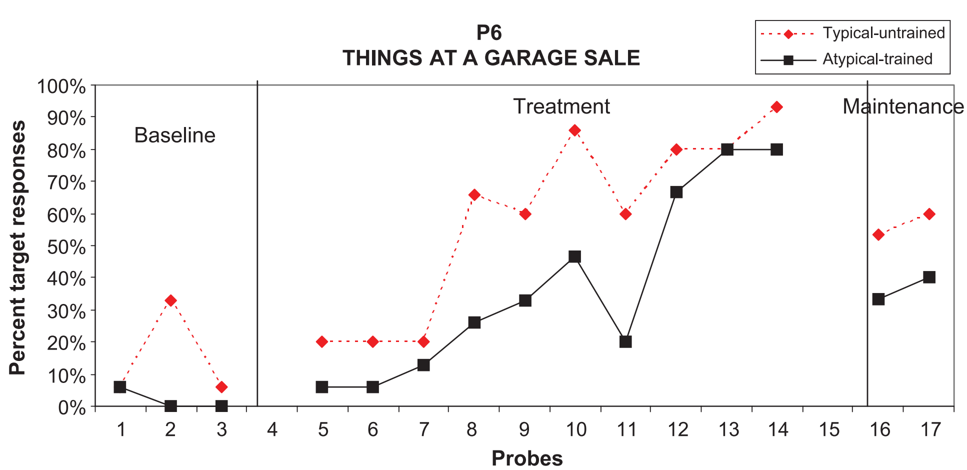
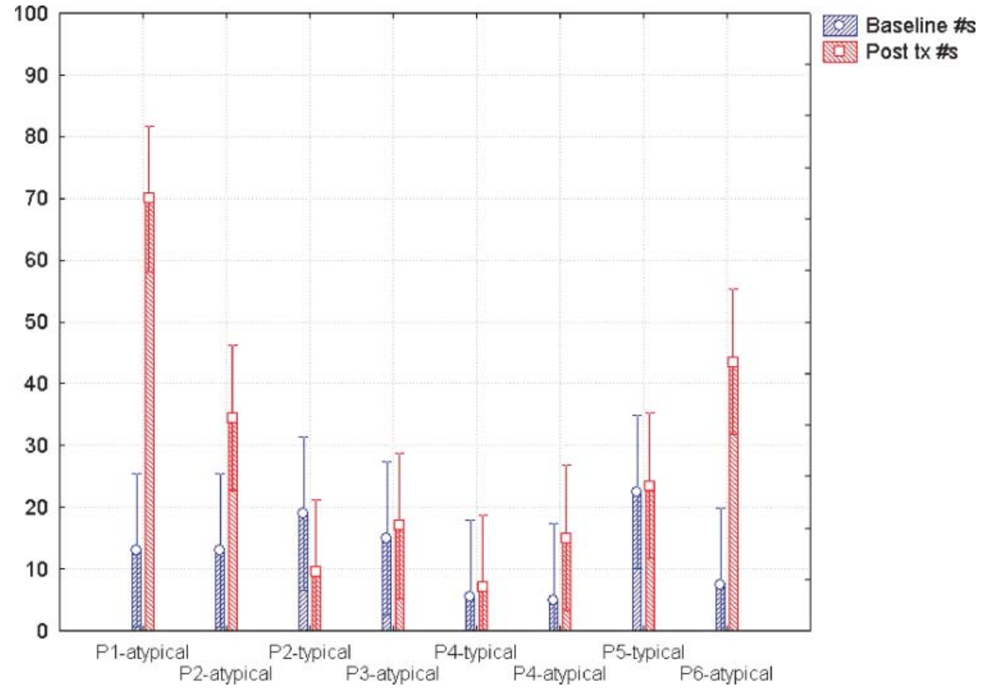


Figure 7. Proportion of responses produced that were tabulated as "other" but that were not target atypical or typical examples. Blue bars indicate the averaged responses generated during the initial two probes, and red bars indicate the average responses generated in the final two treatment (tx) probes.



joining tree cluster analysis (Everitt, Landau, & Leese, 2001). For each category, patients were entered as variables, and the response subtypes were entered as clusters. All variables were equally weighted in the analysis, and distance was computed using a Euclidean measure. To determine distance between clusters, we used the single linkage distance (nearest neighbor) to capture the similarity between neighboring clusters. Therefore, this analysis begins with objects in its own cluster—the nodes representing the distance at which the two closest clusters join—and progresses to a single cluster (agglomerative). Figure 8 illustrates the dendograms for the two categories. On the basis of amalgamation coefficients, the final cluster solution for each category was determined to be cut off at a linking distance of 25 points because the clusters that were merged after a distance of 25 points were quite distinct. Consequently, there are four clusters for *garage sale* (1 = food/electronic/furniture; 2 = kitchen/entertainment/miscellaneous; 3 = home/garden; and 4 = typical examples). Likewise, there were four clusters for *camping* (1 = games/entertainment/wildlife; 2 = other/transportation; 3 = supplies/tools/food/cooking/clothing/personal care; and 4 = production errors). These results indicate that across patients and independent of the typicality of trained stimuli, the evolution of responses produced during the course of treatment clustered along a semantic similarity dimension.

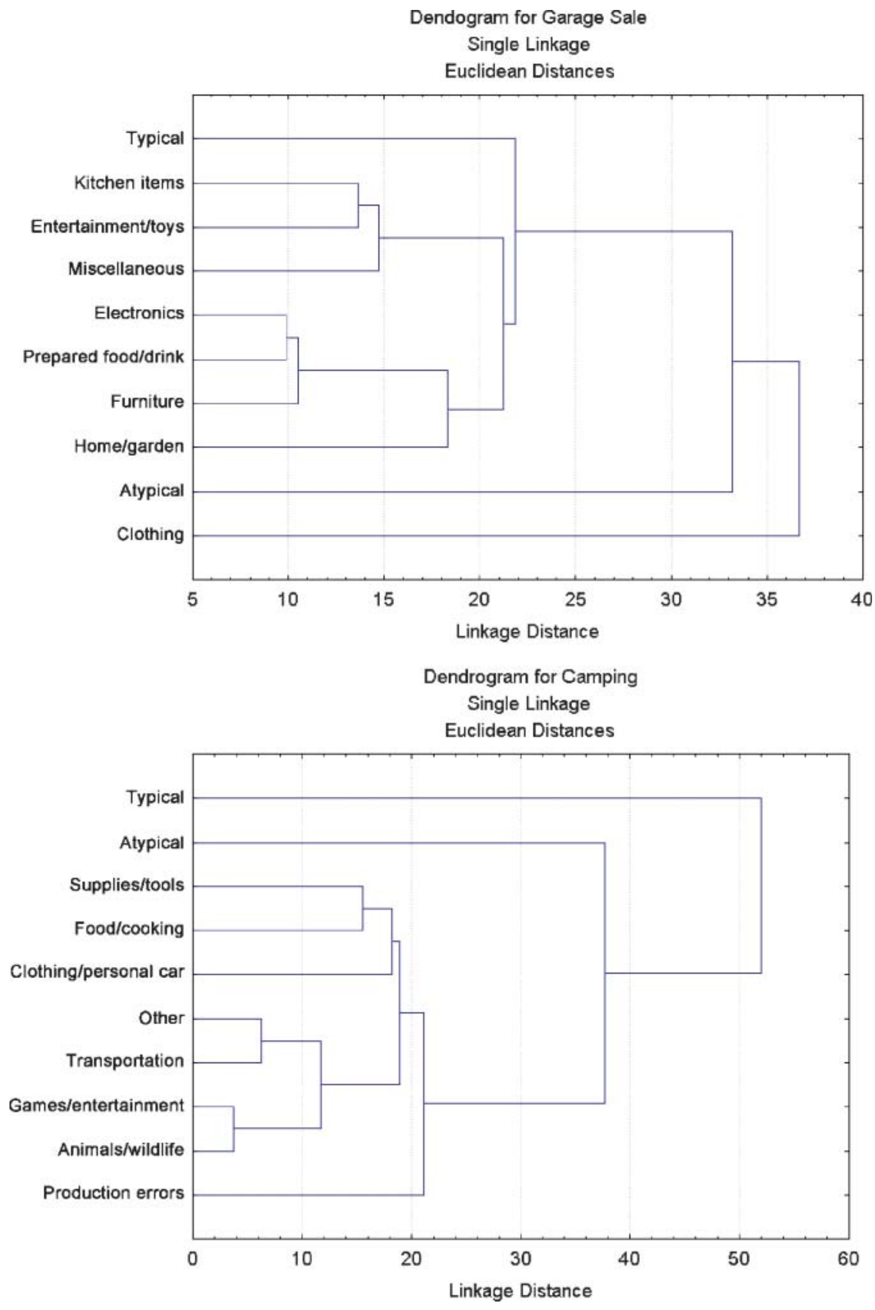
Standardized Tests

Overall, the participants in this study showed improvements or maintained performance on the various standardized measures that were administered pre- and post-treatment. Participants 1, 3, and 4 showed small improvements on WAB Aphasia Quotient (AQ) scores. No patient showed improvements on the BNT, and two of six patients showed improvements on the WAB category scores. On the PALPA and PAPT—with the exception of Writing Picture Names, which improved for all but one participant— improvements on the individual subtests were mixed. Wilcoxon matched pair tests conducted on the pre–post scores did not reveal significance for any subtests.

Discussion

In our previous work (Kiran, 2008; Kiran & Johnson, 2007; Kiran & Thompson, 2003), we showed that training atypical examples is a more efficacious way to facilitate generalization within categories than training typical examples. In the present experiment, we examined the effect of varying exemplar typicality within ad hoc categories in six individuals with aphasia. Recall that in the introduction, we hypothesized that strengthening access

Figure 8. Dendrograms for two categories (*camping, garage sale*) are displayed as horizontal tree clusters. The *x*-axis indicates linkage distance measured as Euclidean distances, and the *y*-axis indicates the cases/clusters considered for each category. See text for further details.



to semantic attributes and phonological representations for target atypical examples would facilitate access to these items as well as to corresponding semantic and phonological representations of untrained typical examples. In contrast, we predicted that strengthening access to semantic attributes and phonological representations of typical items would improve only those items; no generalization to untrained atypical examples was expected. The results of the present experiment confirm our hypotheses and extend our previous findings of typicality treatment for lexical retrieval deficits in two ways. First, the present typicality treatment was effective in improving lexical access within a category generation task, whereas our previous work was focused on improving confrontation picture naming. Second, these results illustrate that typicality within categories can be extended to ad hoc/goal-derived categories, which are categories that have fairly loose category boundaries. On the basis of results shown in Table 4, it is apparent that when atypical members were trained, generation of untrained atypical items improved in three of the five patients (P1, *garage sale*; P3, *garage sale*; and P6, *garage sale*); generation of target typical examples improved in five patients (P1, *garage sale*; P2, *camping*; P3, *garage sale*; P4, *garage sale*; and P6, *garage sale*); and generation of untrained typical items improved in three patients (P1, *garage sale*; P3, *garage sale*; and P4, *garage sale*). In contrast, when typical examples were trained, there was no improvement in generation of untrained typical examples; only one patient showed an improvement (P2, *garage sale*) on the untrained target atypical examples, and no patients improved on generation of untrained nontarget atypical examples. These results suggest that there is benefit to training only atypical examples in the category, as generalization extends to typical examples within the category. A second goal of the study was to closely examine the nature of responses produced in treatment and if the evolution of responses was influenced by whether trained stimuli were typical or atypical items in the category. Results from this analysis suggested that training atypical examples resulted in more examples generated for the category compared with when typical examples were trained.

The results of the present study have clear theoretical and clinical significance. From a clinical standpoint, several studies have aimed at improving lexical access using picture naming as the treatment task by strengthening semantic representations of target items. The use of category fluency as a behavioral variable to improve lexical access has been relatively less common in treatments for patients with aphasia. In the present study, category fluency was considered to be an appropriate task for patients with relatively mild levels of naming impairment. These patients varied on their confrontation picture naming ability, but all were impaired on their ability to retrieve items during a category fluency task. Further, word generation more closely resembles the word-finding required for conversation than confrontation naming, lending itself as a more suitable treatment task for real-world application. Notably, participants who were trained on atypical examples in a category generated more items for the category at the end of treatment as compared with participants who were trained on typical examples of a category. These results supplement our previous work (Kiran et al., 2009) showing that category generation of items can be facilitated through a semantically based treatment.

The theoretical implications of the present study span three domains of semantic representation and are explained below. One obvious theoretical implication is the extension of the typicality effect to ad hoc categories that are not fixed entities in memory but are constructed on a more dynamic basis. As noted in the introduction, ad hoc (or goal-derived) categories (e.g., *a grocery list*) do not have rigid features that constitute category membership; instead, category members follow a loosely combined thread of common features. Although complexity is more difficult to define in such categories, several studies show that typical examples (e.g., *tent*) are more illustrative of the central goal of the category (e.g., *things to take camping*) than are atypical examples (e.g., *playing cards*). The treatment protocol used in this study took into account the fact that core (shared) features in goal-derived categories were restricted to the goal (e.g., *things needed at a camp*); nevertheless, features for typical examples were selected such that they consisted of shared features (e.g., *protection from nature, for sleeping*), whereas atypical examples also included distinctive features that were not shared by other category members (e.g., *things to do at a camp, for personal hygiene*). Consequently, training atypical examples in the category appeared to strengthen semantic features that are relevant for the range of examples within the category, whereas training typical examples improved only the core features in the category. Importantly, training atypical examples resulted in a greater number of overall responses produced than did training typical examples. Further, the cluster analysis for both categories indicated that changes in response types as a function of treatment were along a semantic similarity dimension. Therefore, words in the subcategories that were tokens of *electronics, food/drink,* and *furniture* for *garage sale* evolved similarly as a function of treatment. Likewise, words denoting *supplies, tools, food/cooking,* and *clothing* for *camping* evolved similarly as a function of treatment.

One way to interpret the results is within the now well-established framework of *semantic feature theory* (McRae et al., 1997; McRae, Cree, Seidenberg, & McNorgan, 2005), which states that word meaning is distributed across semantic features. McRae et al. (2005) found that concepts with many shared features are easier to respond to than are concepts with fewer shared features. They further showed that the facilitative effect of the shared features was greater than any inhibitory effect from distinctive features. In another study, using computational simulation and behavioral data (Cree, McNorgan, & McRae, 2006), the same authors showed that distinctive features are activated more quickly when a concept is activated with a subsequent feature name and that distinctive features are better cues for retrieving a concept. Taken together, these studies suggest that both distinctive and shared features influence activation of concepts but in slightly different ways and inform the present study in providing a framework for the generalization patterns observed. Our future work will be focused on disentangling the nature of semantic feature representation that is critical to the prediction of semantic complexity.

The present results also raise an interesting issue regarding semantic distance from the central goal. Although central tendency is not a strong predictor for goal-derived categories (Barsalou, 1985), patients with aphasia tend to be anchored to the center of the referential field for these types of categories (Hough, 1993). This observation may help explain why training items on the periphery of an ad hoc/goal-derived category also improves the items at the center of the semantic field. In other words, all examples selected for treatment fulfilled the goal of the category (e.g., *things to take camping*); however, some examples (typical) had features that were necessary to fulfill the goal, whereas other examples (atypical) were less important in fulfilling the goal but were nonetheless plausible. However distantly linked the atypical items are with the typical items, if the tendency is toward generating the typical items, the slightest overlap in features may trigger activation of typical items and other items that overlap with the goal.

Finally, the results are also consonant with the *situated conceptualization* hypothesis proposed by Barsalou (2003). Barsalou and colleagues (Barsalou, 2003; Yeh & Barsalou, 2006) suggested that concepts activate their relevant associated background situations and allow a perceptual experience of the person imagining the event or concept. In effect, the person simulates a perceptual, motor, and introspective experience whenever presented with a concept or an event. Inferences regarding concepts are generated in a situation-specific manner, and when individuals are asked to retrieve instances for a concept, they likely imagine themselves in the specific situation (Vallee-Tourangeau, Anthony, & Austin, 1998; Yeh & Barsalou, 2006). Applying this notion to the present study, we can posit that when patients are being trained on a specific concept (e.g., *things at a garage sale*), they simulate themselves being in that specific situation and generate inferences about objects and events likely to be present in that situation (Barsalou, 2003). Therefore, training the atypical examples likely requires patients to imagine themselves within a context that includes images of a larger set of examples than when patients are trained on the typical examples, which requires activation/imagination of only those items that directly fulfill the goal.

Results from the standardized assessments administered pre- and post-treatment indicate that participants showed modest improvements or maintained performance. P1, P3, and P4 showed small but significant improvements on WAB AQ scores. These findings are consistent with a previous study wherein we argued that patients who received a semantically based naming treatment improved on standardized language tests of aphasia as compared with a nontreated patient group (Warfield & Kiran, 2008). In that study, we found that patients who received a semantically based naming treatment significantly improved by an average of 8.6 points on the WAB AQ at the end of the maintenance period. In contrast, the nontreated group decreased an average of 5.7 points on the WAB AQ; however, this decrease was not significant. Results on other standardized tests (BNT, PALPA, and PAPT) were inconclusive. One possible reason for this lack of significance is that participants P1, P2, P3, and P6 were already at ceiling level on most subtests at pretesting.

Conclusion

In conclusion, this study has added to the corpus of research showing the effectiveness of using complexity in treatments for word retrieval. We should note that although the experimental design is set up as a counterbalanced design, because of the length of treatment (20 weeks) and the relatively rigid structure of treatment, not all participants completed the entire treatment protocol. Therefore, there is an imbalance in the number of participants who received typical treatments versus atypical treatments. To mitigate this potential confound, we set a more stringent criteria for successful acquisition and generalization. Nevertheless, the results need to be interpreted in the context of a limited number of participants. The results of the present study can be considered as a preliminary proof-of-concept for the use of a typicality treatment within ad hoc categories using a task such as category generation. We intend to continue examining the effects of typicality training on a larger and more diverse sample of aphasia patients as well as with different types of ad hoc categories.

Acknowledgments

This research was supported by National Institute on Deafness and Other Communication Disorders Grant DC006359-03 and a New Century Research Scholars Grant from the American Speech-Language-Hearing Foundation to the first author. We thank Lisa Edmonds and Shilpa Shamapant for their assistance during various stages of the project and Crystal Cavin and Jessica Guzman for their role in analyzing the naming errors. Finally, we thank the participants in the experiment for their patience and cooperation.

References

Barsalou, L. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes, 18,* 513–562.

Barsalou, L. W. (1983). Ad-hoc categories. *Memory and Cognition, 11,* 211–227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 629–654.

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychological Review, 16,* 161–169.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A,* 497–505.

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 643–658.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). London, England: Arnold Publishers.

Goodglass, H., Kaplan, E., & Weintraub, S. (1983). *Boston Naming Test*. Philadelphia, PA: Lea & Febiger.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18,* 441–461.

Helm-Estabrooks, N. (2001). *Cognitive Linguistic Quick Test*. London, England: Harcourt Assessment.

Hough, M. S. (1993). Categorization in aphasia: Access and organization of goal-derived and common categories. *Aphasiology, 7,* 335–357.

Hough, M. S. (2007). Incidence of word finding deficits in normal aging. *Folia Phoniatrica et Logopaedia, 59,* 10–19.

Howard, D., & Patterson, K. (1992). *Pyramids and Palm Trees*. London, England: Harcourt Assessment.

Kay, J., Lesser, R., & Coltheart, M. (1992). *The Psycholinguistic Assessment of Language Processing in Aphasia*. Hove, United Kingdom: Erlbaum.

Kertesz, A. (1982). *The Western Aphasia Battery*. Philadelphia, PA: Grune and Stratton.

Kiran, S. (2007). Semantic complexity in the treatment of naming deficits. *American Journal of Speech-Language Pathology, 16,* 1–12.

Kiran, S. (2008). Typicality of inanimate category exemplars in aphasia treatment: further evidence for semantic complexity. *Journal of Speech, Language, and Hearing Research, 51,* 1550–1568.

Kiran, S., & Bassetto, G. (2008). Evaluating the effectiveness of semantic based treatment for naming deficits in aphasia: What works? *Seminars in Speech and Language, 29,* 71–82.

Kiran, S., & Johnson, L. (2008). Semantic complexity in treatment of naming deficits in aphasia: Evidence from well-defined categories. *American Journal of Speech-Language Pathology, 17,* 389–400.

Kiran, S., Sandberg, C., & Abbott, K. (2009). Treatment for lexical retrieval using abstract and concrete words in persons with aphasia: Effect of complexity. *Aphasiology, 23,* 835–853.

Kiran, S., & Thompson, C. K. (2003). The role of semantic complexity in treatment of naming deficits: Training semantic categories in fluent aphasia by controlling exemplar typicality. *Journal of Speech, Language, and Hearing Research, 46,* 773–787.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations in word meaning. *Journal of Experimental Psychology: General, 126,* 99–130.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, 37,* 547–559.

Pexman, P. M., Holyk, G. G., & Monfils, M.-H. (2003). Number-of-features effects and semantic processing. *Memory and Cognition, 31,* 842–855.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111–144). New York, NY: Academic Press.

Rosch, E. (1975). The nature of mental codes for color categories. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 303–322.

Sandberg, C., Sebastian, R., & Kiran, S. (2011). *Processing of typicality within well defined categories and ad-hoc categories in patients with aphasia*. Manuscript submitted for publication.

Thompson, C. K. (2007). Complexity in language learning and treatment. *American Journal of Speech-Language Pathology, 16,* 3–5.

Vallee-Tourangeau, F., Anthony, S. H., & Austin, N. G. (1998). Strategies for generating multiple instances of common and ad-hoc categories. *Memory, 6,* 555–592.

Vigliocco, G., Vinson, D. P., Damian, M. F., & Levelt, W. (2002). Semantic distance effects on object and action naming. *Cognition, 85,* B61–B69.

Warfield, E., & Kiran, S. (2008, November). *Impact of model-based semantic treatment on standardized tests of aphasia*. Paper presented at the American Speech-Language-Hearing Association Convention, Chicago, IL.

Yeh, W., & Barsalou, L. W. (2006). The situated nature of concepts. *American Journal of Psychology, 119,* 349–384.

Appendix A. Target and distractor semantic features for the two categories.

Type of semantic feature	Garage sale	Type of semantic feature	Camping
Target	Entertainment	Target	Safety items
Target	Furniture	Target	Needed for cooking
Target	To sleep on	Target	For cleaning
Target	Kitchen or cooking items	Target	For personal hygiene
Target	Art	Target	For comfort
Target	Bedroom items	Target	For warmth
Target	Men's clothing	Target	Things needed at a camp
Target	Clothing	Target	Multipurpose
Target	Things that cover the body	Target	For repairs
Target	Electronics	Target	Things to do at a camp
Target	Items for travel	Target	To take out what you take with you
Target	Baby items	Target	For carrying/storing things
Target	Unwanted/unneeded	Target	For lighting at night
Target	Outdated/outgrown	Target	Protection from nature (not weather)
Target	Collectible	Target	Protection from weather
Target	Estate sale	Target	Clothing
Target	For kids	Target	For setting up camp
Target	Household items	Target	For sleeping
Target	Exercise/recreation	Target	Picnic goods
Target	Memorabilia	Target	Obtain at sporting store
Target	Communication	Target	Buy at drugstore
Target	Ways to hear music	Target	Optional items
Target	For research	Distractor	Found at a crime scene
Target	Things for reading	Distractor	Found under the soil
Target	Things that provide light	Distractor	Carries disease
Target	Decorations	Distractor	Uses bullets
Target	Office- or work-related items	Distractor	Kills for food
Target	Useful items	Distractor	Makes music

Appendix B. Treatment protocol.

One set of items ($N = 15$; e.g., typical or atypical) was treated at a time. For each item in the training set, the treatment protocol was as follows:

1. Category generation. This step was performed only once at the beginning of each session. The patient was asked to generate as many examples as possible for the *things at a garage sale* category. The clinician wrote down all the responses produced and displayed them to the participant as feedback. Then, the clinician selected one item from the target list (typical or atypical) and said "*Okay, let's pick this word (bicycle). Let's go through the training steps, and I'll help you understand more about the features/details of (bicycle) and why it's something you can have at a garage sale.*"
2. Category sorting. This step was performed once at the beginning of each session. The clinician placed written category cards on the table: *Things at a garage sale* and *things that you see during Christmas time (distractor category).* The clinician presented the patient with the word cards and asked him or her to sort the cards according to their superordinate category by placing the cards on the category cards. If the patient categorized a word incorrectly, he or she was given feedback: "*Are you sure this (bicycle) is found at a crime scene? It's actually something you can sell at a garage sale.*"
3. Feature generation/selection. The clinician placed the target word at the center of the table and asked the patient to generate as many attributes as possible (at least 4–6) regarding the target (e.g., *bicycle*) that make it a good item to fit into the category (e.g., *things at a garage sale*). For example, the participant could say, "*children outgrow it, old model,*" and so forth. Then, the clinician presented the patient with the features of the target category and asked the patient to select the first six semantic features that were pertinent to *bicycle*. For example, for *bicycle,* the features that were practiced included five that were pertinent to the example, (e.g., *outgrown/outdated*), five that belonged to the category but not the example (e.g., *bedroom*), and five that did not belong to the category (e.g., *lives in a forest*). The clinician selected six features and read the features aloud to the patient.
4. Yes/no questions. The clinician then removed the target picture and the written phrases and instructed the patient "*I'm going to ask you some questions about (bicycle) now. Please answer yes or no for each of these questions.*" The clinician asked a total of 15 questions, up to five questions that were relevant to the target example, (e.g., *has pedals*), five that belonged to the category but not to the example (e.g., *is feminine*), and five that did not belong to the category (e.g., *lives in a forest*).