Memory & Cognition
2003, 31 (5), 673-689

ERP and behavioral evidence of individual differences in metaphor comprehension

VICTORIA A. KAZMERSKI, DAWN G. BLASKO, and BANCHIAMLACK G. DESSALEGN
Penn State Erie, The Behrend College, Erie, Pennsylvania

In two experiments, we examined individual differences in metaphor processing. In Experiment 1, the subjects judged the literal truth of literal, metaphorical, and scrambled sentences. Overall, metaphors were more difficult to judge as false, in comparison with scrambled controls, suggesting that the metaphorical meaning was being processed automatically. However, there were individual differences in that high-IQ subjects showed more interference. These effects were reflected in ERP amplitude differences at the onset of N400 and after the response. In Experiment 2, the subjects completed IQ tests and a series of working memory tests and then rated and interpreted the same set of metaphors. The results showed that IQ was correlated with working memory capacity and that low-IQ subjects had similar ratings but poorer quality interpretations than did high-IQ subjects. The results were most consistent with a constraint satisfaction approach to metaphor comprehension.

Metaphors abound in conversational language. We can describe a politician as a snake and a lawyer as a shark without even realizing that a conceptual leap between the metaphor's *topic* (politician) and its *vehicle* (snake) has been made. The very fact that many metaphors seem to be produced and understood effortlessly has been of great interest to researchers. The issue of the exact time course of figurative meaning activation has been key to the controversy over whether figurative language is directly or indirectly processed (Glucksberg, 2001). According to the classic indirect model (e.g., Searle, 1979), when a metaphor is first encountered, the first stage of processing is to attempt a literal interpretation. If a literal interpretation is impossible or if it does not make sense in that particular context, a special figurative processing system takes over and uses pragmatic information to infer a figurative meaning. The model made several testable predictions. The first is that understanding a metaphor should take longer than understanding a literal paraphrase, because literal processing must always occur first. The second prediction is that metaphor interpretation should be a secondary optional process. In order to test the second prediction, Glucksberg and colleagues (Glucksberg, Gildea, & Bookin, 1982) developed the metaphor interference task.

In the metaphor interference task, subjects are asked to read a series of word strings and judge whether or not they are literally true. The critical conditions include literally true sentences (*The robin is a bird*), literally untrue but metaphorically true sentences (*The divorce is a nightmare*), and scrambled sentences made from the same topic and vehicle terms in such a way as to make them less sensible both literally and metaphorically (*The divorce is a table*). The logic of the task was that if readers were asked to judge the literal truth of metaphors, the decision to judge them as untrue could be made at the first, literal stage, thereby making the additional processing of any metaphorical meaning unnecessary. However, data from a large number of studies showed that subjects often took longer to judge metaphors as literally untrue than to reject the less meaningful scrambled sentences (Glucksberg et al., 1982; Light, Owens, Mahoney, & La Voie, 1993; Wolff & Gentner, 2000). Therefore, the metaphor interference task was interpreted to reflect the automatic and obligatory nature of metaphor processing (Glucksberg & Keysar, 1990). That is, even when explicitly told not to, people simply could not help but process the meaning of metaphors. When these findings were combined with evidence from eye-tracking studies that metaphors did not always take longer to read than literal sentences (Inhoff, Lima, & Carroll, 1984) and cross-modal priming evidence of immediate figurative meaning activation (Blasko & Connine, 1993), the field was led away from the literal-first indirect models and toward models suggesting that figurative meaning could be computed directly from the information available in the communicative environment (e.g., Gibbs, 1994; Glucksberg & Keysar, 1990).

Recent work has shown that this issue was more complicated than was originally thought. First, there are times when understanding figurative language is clearly a less direct and more time-consuming process than literal understanding. For example, proverbs consistently take longer to comprehend than literal paraphrases (Honeck,

This study was supported by Penn State Erie Undergraduate Student Research grants. We thank Athena Farantzos, Shannon Lenze, Jessica Turos, Amanda Ervin, Andrea Furman, Jacelyn Tetuan, Alejandra Marroquin, and Eden Roseborough for assisting in collecting and analyzing data. Correspondence concerning this article should be addressed to V. A. Kazmerski, Humanities and Social Sciences, Penn State Erie, 5091 Station Road, Erie, PA 16563-1501 (e-mail: vak1@psu.edu).

Welge, & Temple, 1998). Second, many have challenged the view that figurative and literal language are really distinct and, instead, have considered them to exist on a continuum of conventionality. Finally, characteristics of a metaphor, such as familiarity and aptness, have been shown to influence the time course of meaning activation. For example, readers show both shorter gaze durations (Blasko & Briihl, 1997) and earlier activation of figurative meaning to familiar and apt metaphors, in comparison with unfamiliar and less apt metaphors (Blasko & Connine, 1993). The metaphor interference task has been widely used in metaphor research because it appears to be sensitive to characteristics of the metaphor and its surrounding context. Gregory and Mergler (1990) did not find evidence of the metaphor interference effect (MIE) and hypothesized that this was because they had used metaphors that were less apt and/or less familiar than the original set. Less apt metaphors, created by using the quantifier *all* (e.g., *All marriages are iceboxes*), showed no MIE without supportive context (Gildea & Glucksberg, 1983). Thus, the metaphor interference task is a widely used measure of the earliest and most automatic processing of metaphor (Gibbs, 1994). In the present paper, we will examine it further in order to determine its time course and to investigate whether it is susceptible to individual differences.

Individual Differences in Metaphor Processing

Most models of metaphor processing make the simplifying assumption that all adult readers are essentially the same. Yet any professor of literature will tell you that a meaningful metaphor to one student may be completely obtuse to another. Take, for example, the literary metaphor spoken by Buckingham in Shakespeare's *Life of King Henry the Eighth*, "No man's pie is freed / From his ambitious finger." Some students recognize the idiom "a finger in every pie"[1] and, from the context of the plot, recognize that this might apply to those who meddle in the affairs of others. However, other students see little or no meaning and continue to have difficulty with such metaphors even after careful explanation.

Despite the fact that metaphor comprehension has historically been equated with intellectual ability and higher level cognition, there has been relatively little research investigating exactly why such wide variations in the understanding of metaphors occur. In one of the few studies to date, Trick and Katz (1986) found a relationship between analogical reasoning scores and ratings of metaphor aptness. Those with higher analogic reasoning scores gave higher ratings to those metaphors whose terms came from dissimilar domains. They also appeared to be more sensitive to the precise structural correspondence between domains. According to Gentner and colleagues, the structural alignment and mapping between topic and vehicle domains lies at the heart of metaphorical mapping (Gentner & Wolff, 1997). Therefore, one source of individual differences may lie in a person's skill in bridging the semantic domains of topic and vehicle. They must activate features of the vehicle that fit the constraints of the topic, while at the same time filtering out or suppressing the irrelevant features (Gernsbacher & Keysar, 1995). In some cases, this involves creating entirely new emergent features that were not salient in either topic or vehicle domains, as well as using inference processes to extract the correct structural dimensions of comparison from the context.

Individual differences in verbal IQ have often been explained, in part, as arising from constraints of general cognitive capacity. Working memory for language, as indexed by complex span tasks, has been shown to predict individual differences in both auditory (Connine, Blasko, & Wang, 1994) and visual language tasks (Gernsbacher & Faust, 1991; Just & Carpenter, 1992). The capacity theory of language (Just & Carpenter, 1992) suggests that differences in working memory for language provide functional limits on the speed and efficiency of comprehension. Because working memory capacity is hypothesized to involve both the processing and the storage of the partial products of comprehension, capacity-based performance decrements would be predicted in cases in which an individual's resources are exceeded because of inherent capacity limitations, poor efficiency of the particular processes involved in the task, or both (Just & Carpenter, 1992).

Taken as a whole, this research suggests that for any given metaphor, comprehension difficulty may be a function of a multitude of factors, including the distance of the semantic mapping, the demands of the task, and the verbal ability of the individual. For example, if subjects are given unlimited time to rate easy metaphors for comprehensibility, this should require relatively few resources, and therefore, there should be few individual differences. However, if the task is speeded or if it involves not just judging *whether* an utterance makes sense but also explaining *how* it makes sense, considerable resources must be marshaled, and those with lower intellectual capacity may have more difficulty. Most current models of metaphor comprehension make little mention of this tradeoff. In fact, research has been conducted with a focus on stimulus characteristics while, for the most part, ignoring subject characteristics.

One of the few models of metaphor comprehension that seems amenable to the interaction of stimulus and subject characteristics is the predication approach recently advanced by Kintsch (2000, 2001). According to this constraint satisfaction approach, stimuli and subject characteristics are expected to flexibly interact. The linguistic experience of the individual will, over time, set up a semantic network structure that will then flexibly interact with contextual information and task demands to dynamically compute meaning on line for each utterance, whether metaphorical or literal. However, this predication approach is a computational model; therefore, it does not provide information on how these processes occur in the brain.

Event-Related Brain Potentials and Language Comprehension

Despite a recent upsurge of interest in cognitive neuroscience, relatively few studies have directly addressed the issue of how figurative language is processed in the

healthy brain. The majority of work has focused on clinical populations and has used off-line measures of interpretation. For example, a large number of studies have shown that patients with right-hemisphere damage may have difficulty dealing with figurative language, such as metaphors (Winner & Gardner, 1977), indirect requests (Hirst, LeDoux, & Stein, 1984), and jokes (Bihrle, Brownell, & Powelson, 1986). Research on the normal brain has tended to focus on the relative contribution of the two hemispheres and has suggested that the right hemisphere might be somewhat better prepared to deal with the peripheral meanings often needed in metaphor processing (Beeman, 1993; Beeman et al., 1994; Burgess & Chiarello, 1996; Titone, 1998). Recent research in which PET (Bottini et al., 1994) was used showed greater right-hemisphere activation, especially in the right prefrontal cortex, the middle temporal gyrus, the precuneus, and the posterior cingulate, when subjects read metaphors, in comparison with literal sentences.

Event-related brain potentials (ERPs) are a technique well suited to explore both the time course of the MIE and potential individual differences. ERPs are time-locked recordings of voltage changes within the brain that are recorded from the scalp in response to sensory, motor, or cognitive events. They provide unsurpassed temporal resolution, making them ideal for the study of the time course of sentence processing (Fabiani, Gratton, & Coles, 2000; Kutas, 1997; Kutas, Federmeier, Coulson, King, & Münte, 2000). ERPs have the additional advantage of allowing random presentation of stimuli across conditions. In addition, through the use of multiple scalp electrode sites, inferences can be made about the localization of the source of activation (R. Johnson, 1993). Although such inferences are limited when compared with the capabilities of fMRI or PET (because scalp-recorded potentials are summed from a large populations of neurons), they can provide valuable information about basic comparisons between right and left hemispheres and differential activity in the front and the back of the brain.

ERPs have been shown to be sensitive to individual differences in IQ (Stelmack & Houlihan, 1995) and reading ability (Lovrich, Cheng, Velting, & Kazmerski, 1997) and in clinical populations (e.g., Kazmerski & Friedman, 1998). One ERP component that has been widely used in studies of language processing is the N400 component. This negative voltage deflection occurs around 400 msec after stimulus onset and was originally found when sentences with and without semantic anomalies were compared. However, it has also been found to be larger in cases in which semantic integration is more difficult and, therefore, is related to cloze probabilities (Coulson, 2001; Kutas, 1997; Kutas et al., 2000). A second component that has been identified in language studies is a positive-going wave that usually begins 500–600 msec after stimulus presentation. This component, the P600, has been more closely associated with syntactic anomalies (Osterhout & Holcomb, 1992) but has also been shown to index social factors, such as gender stereotypical language use (Osterhout, Bersick, & McLaughlin, 1997). A late positivity has also been shown to index successful comprehension. For example, subjects who understood the punch lines of jokes showed larger late positivity than did those who had not (Coulson, 2001).

An assumption of these techniques is that the subject is actively evaluating the stimulus and that the corresponding measures are reflecting that evaluation. However, when looking at the MIE, we must also consider that the subject must select the response appropriate to the sentence and possibly, in the case of a metaphor, inhibit or suppress an automatically activated meaning in order to respond that a statement is not meaningful. ERPs can be helpful in dissociating such response-related activity. For example, the lateralized readiness potential is known to reflect motor preparation for a response (Coles, Smid, Scheffers, & Otten, 1995).

There has been relatively little attention paid to nonliteral language in the ERP literature. In one exception (Pynte, Besson, Robichon, & Poli, 1996), ERPs were recorded as subjects read familiar metaphors (*Those fighters are lions*), unfamiliar metaphors (*Those apprentices are lions*), or literal control sentences (*Those animals are lions*). The vehicle of the metaphor showed a larger N400 component than did the last word of the literal sentences. They also found that supportive context reduced the magnitude of the N400 but that there was no evidence of differences later in processing (1,000–1,400 msec) and no evidence of laterality differences that could be attributed to metaphorical processing.

The Present Study

There have been a few studies that have shown that the quality of off-line metaphor interpretations was correlated with general cognitive measures (e.g., working memory for language; verbal SAT scores, Blasko, 1999; mental capacity, J. Johnson & Pascual-Leone, 1989). However, there has been little work in which researchers have looked at potential individual differences at the earliest and most automatic level of metaphor processing, when relevant features are first activated and irrelevant ones inhibited. The MIE provides just such an opportunity. Therefore, in the present study we investigated individual differences in the processing of the same set of metaphors in both an on-line task and an off-line task. In Experiment 1, we used ERPs to map the size and time course of the MIE in relation to IQ. In Experiment 2, we collected ratings and interpretations of the same metaphors, as well as working memory and IQ scores from our subjects.

One purpose of the present study was to replicate the MIE and to investigate potential individual differences by partitioning the sample on the basis of IQ scores from a standardized test. IQ scores were selected as the measure of individual differences, since they are widely used, have well-known psychometric properties, and are easily administered. To make the task more sensitive to capacity limitations and more compatible with standard ERP methodology, we modified the full sentence pre-

sentation of the typical MIE experiment to a rapid serial visual presentation (RSVP). We examined processing between three key conditions: literal, metaphorical, and scrambled. Subjects were asked to judge whether each sentence was literally true; thus, in the case of the metaphorical condition, if the metaphorical meaning is automatically activated, it would need to be inhibited in order to make the response. The ERP indices of the metaphor interference task would be seen in a divergence of the average waveforms for the metaphorical and the scrambled conditions. In order to reduce individual differences based on basic vocabulary or general knowledge, we extensively normed our stimuli and avoided more difficult and obtuse literary metaphors, as well as highly conventionalized/lexicalized metaphors, and instead focused on a set of metaphors with a moderate range of familiarity and aptness.

Most methods of studying on-line language comprehension can look only at a single snapshot in time—that is, reaction time (RT) differences shown in the MIE tell us only that, at the point at which the behavioral response was made, there was some average difference between conditions. ERPs go beyond this and can provide information about processing before and after a response. It has been argued (Wolff & Gentner, 2000) that the relatively fast RTs of the MIE (in the order of 1,200 msec), in comparison with the much lengthier times found in interpretation tasks (5 sec or more), reflect a very early stage of comprehension, before any later interpretation or reflection can occur. The question is, at what point does figurative activation occur. To conduct an analysis of the time course of processing, we examined the three conditions across 200-msec latency bins. We expected to see a larger N400 to the metaphorical and scrambled conditions than to the literal condition. One of the most interesting questions theoretically is the mapping of the time course of the MIE. If there are differences between the metaphorical and the scrambled conditions, at what point will they emerge and for how long will they remain? To the extent that the MIE indexes automatic meaning activation, we should see a difference between the metaphorical and the scrambled conditions begin to emerge in the earliest bin, 200–400 msec, which includes the early part of the N400 component, so that N400 amplitude should be greater for the scrambled condition than for the metaphorical condition.

Our final question was whether there would be evidence for individual differences in the MIE. Some research suggests that high-skilled readers have more efficient suppression mechanisms (Gernsbacher & Faust, 1991). If this is the case, the high-IQ group might be better at the inhibition task and show less difference between metaphorical and scrambled conditions. However, on the basis of the results of work showing that high comprehenders' sensitivity to syntactic constraints leads to a time-consuming garden path (King & Just, 1991), it is more likely that the high-IQ group will find metaphor processing both easier and more automatic and, therefore, will have more difficulty ignoring metaphorical meaning. If so, activation of the metaphorical meaning may be seen as greater positivity for the high-IQ group.

EXPERIMENT 1

Method

Subjects. Forty-eight (50% female) undergraduates participated in the study. The majority received class credit for participation; the others were paid $20 for participating. The mean age was 19.5 years (SD = 2.8). These subjects were recruited from a larger pool of over 100 students on the basis of their performance on the Kaufman Brief Intelligence Scale (KBIT; Kaufman & Kaufman, 1990). Each of the three groups had 16 subjects. The high-IQ group had IQs of 115 or above (M = 120.8, SD = 4.2) . The medium-IQ group had IQs of 100–114 (M = 105.9, SD = 3.7). The low-IQ group had IQs less than 100 (M = 95.3, SD = 4.7). The subjects were all right-handed and were native English speakers. They reported no serious visual or reading disabilities and no psychological or physical health problems.

Materials. Forty metaphors were chosen as the base stimuli for the experiment. Table 1 shows example sentences.[2] Two lists were constructed so that 20 of the metaphors were presented in their complete form in each list—for example, *The cigar is a skunk*. The topics and vehicles of the other 20 metaphors were scrambled in such a way as to attempt to make them less meaningful—for example, *The cigars are roller coasters*. The metaphorical and scrambled sentences did not differ on cloze probability.[3] Forty literal sentences were developed in order to balance the number of true responses and were used in both lists. Each sentence contained five words.

ERP recordings. EEG was recorded from midline (Fz, Cz, and Pz) and five lateral pairs of scalp sites (F3/4, C3/4, P3/4, T5/6, and O1/2), employing standard 10–20 placements, all referred to nose tip, using an Electro-cap. Vertical (supraorbital to infraorbital bipolar recording) and horizontal (outer canthus of each eye) eye movements were also recorded. ERPs were recorded using a 16-channel amplifier system from Contact Precision Instruments with a 0.01- to 30-Hz bandpass and were sampled at the rate of 200 Hz. The EEG was digitized on line and stored as single trials with corresponding behavioral responses, using the PC-EXP program (Gratton, 1997). The ERPs were averaged off line according to condition and stimulus type. Trials containing eye artifacts were corrected off line, using the technique described by Gratton, Coles, and Donchin

Table 1
Examples of Metaphorical, Scrambled, and Literal Sentences

Sentence Type		
Metaphorical	Scrambled	Literal
The beaver is a lumberjack.	The rumor was a lumberjack.	Tulips grow from a bulb.
The ant trails were freeways.	The sermon is a freeway.	A macaw is a parrot.
Sermons can be sleeping pills.	Swimmers can be sleeping pills.	The guppy is a fish.
The cigar is a skunk.	The cigar is a roller coaster.	Cotton is a natural fabric.
The family was a fortress.	The ant was a fortress.	The hammer is a tool.

(1983). ERPs were recorded to the final word of each sentence so that, across lists and conditions (metaphorical or scrambled), average recordings reflected the same final words. For the metaphors, this corresponded to the metaphorical vehicle. The reported averages reflect correct responses.

Individual difference measures. Standardized tests of intelligence and reading skills[4] were administered to the subjects as the basis for grouping and to screen for potential reading difficulties that may have interfered with the completion of the task. The KBIT (Kaufman & Kaufman, 1990) is a brief intelligence scale that is individually administered in approximately 30 min. It includes verbal and matrices subscales and a composite IQ score. It correlates with the WAIS–R Full Scale IQ ($r = .75$, $p < .001$).

Procedure. The subjects were run in individual sessions lasting approximately 2 h. The individual difference measures were administered first, followed by the ERP recording session. The ERP session included four blocks of a 50-trial oddball task prior to the metaphor task.

In the metaphor task, the sentences were presented one word at a time for a duration of 300 msec, with an interstimulus interval of 500 and 2,000 msec between sentences. The last word of the sentence was followed by a period that indicated to the subject that a response was required. The subjects were told, "You will be presented with a series of sentences that will appear one word at a time on the screen in front of you. When you see the period at the end of a word you will know that the sentence is over. At that point, decide as quickly as possible whether the sentence is literally true." They were then asked to press either the right or the left shift key on a computer keyboard, using the corresponding hand. Hand of response was counterbalanced across subjects. The subjects first completed a set of 10 practice trials that included metaphorical, scrambled, and literal sentences. All the subjects reached at least 80% accuracy on the practice trials.

Results

Behavioral analyses. Table 2 shows the means and standard deviations for mean RTs and accuracies across groups. Responses less than 100 msec and greater than 1,500 msec were discarded as outliers. This accounted for less than 3% of the data and was similar across conditions. All data were analyzed with both subject (F_s) and item (F_i) analyses.

Because the MIE is seen primarily in judgment time, accuracies were analyzed primarily to rule out the possibility of a speed–accuracy tradeoff. As can be seen in Table 2, the subjects were quite accurate at the literal judgment task, averaging 96% correct across conditions. Accuracies were slightly higher in the metaphorical and scrambled conditions, which did not differ from each other, than in the literal condition. This resulted in a main effect of sentence type that was significant in the subject analysis, but not in the item analysis [$F_s(2,90) = 6.54$, $MS_e = 18.91$, $p = .002$; $F_i(2,234) = 0.524$, $MS_e = 175.75$, $p = .59$]. Neither the effect of IQ group nor the interaction of sentence type with IQ group was significant [IQ, $F_s(2,45) = 1.58$, $MS_e = 36.35$, $p = .22$, and $F_i(2,117) = 0.61$, $MS_e = 299.79$, $p = .55$; interaction, $F_s(4,90) = 0.82$, $MS_e = 18.96$, $p = .51$, and $F_i(4,234) = 0.23$, $MS_e = 175.75$, $p = .92$].

Correct RTs were analyzed with 3 × 3 (sentence type [metaphorical, scrambled, or literal] × IQ group [low, medium, or high]) mixed analyses of variance (ANOVAs), with sentence type as a within-subjects variable and IQ as a between-subjects variable. The results showed a significant main effect of sentence type in both subject and item analyses [$F_s(2,90) = 37.06$, $MS_e = 2{,}010.32$, $p < .001$; $F_i(2,234) = 23.46$, $MS_e = 6{,}934$, $p < .001$]. The main effect of IQ group was significant in the item analysis and approached significance in the subject analysis [$F_s(2,45) = 3.02$, $MS_e = 37{,}746.78$, $p = .059$; $F_i(2,117) = 41.55$, $MS_e = 6{,}814.95$, $p < .001$]. The interaction did not reach significance in either analysis [$F_s(4,90) = 0.97$, $MS_e = 2{,}010.32$, $p = .42$; $F_i(4,234) = 0.699$, $MS_e = 6{,}934.00$, $p = .60$]. These results tell us primarily that the true decision was faster than the false one and that IQ moderated speed of processing.

The more interesting question was whether there was an MIE (metaphor − scrambled) evident for each of the three individual groups. Therefore, a second series of ANOVAs was conducted in which only the interaction of the critical sentence types (metaphorical or scrambled) with IQ group (low, medium, or high) was examined. The results showed that even when only sentences that were judged as being not literally true were examined, there was still a significant effect of sentence type in both the subject and the item analyses [$F_s(1,45) = 8.26$, $MS_e = 1{,}625$, $p < .01$; $F_i(1,117) = 4.48$, $MS_e = 5{,}597$, $p < .05$]. This replicates the basic MIE—that is, when collapsed across IQ group, the metaphors were rejected significantly more slowly than the scrambled sentences. The main effect of IQ group was also significant in both subject and item analyses, with high-IQ subjects tending to be faster [$F_s(2,45) = 3.32$, $MS_e = 23{,}696$, $p < .05$; $F_i(1,117) = 31.03$, $MS_e = 5{,}653$, $p < .001$]. This finding was followed up by a series of planned comparisons in which the MIE (metaphor − scrambled) was examined separately for the three groups. Family-wise error was controlled using the Fisher LSD procedure. Inspection of the group means shows a linear increase in the size of the effect across groups. The degree of interference was only 11 msec for the low-IQ group, 27 msec for the medium-IQ group, and 35 msec for the high-IQ group.

Table 2
Mean Reaction Time (in Milliseconds) and Accuracy of the Sentences by IQ Group

	Reaction Time		Accuracy	
Sentence Type	*M*	*SD*	*M*	*SD*
	High-IQ Group			
Metaphorical	747	112	97.5	4.1
Scrambled	712	117	96.7	5.7
Literal	660	133	95.2	5.1
	Medium-IQ Group			
Metaphorical	796	102	97.0	3.9
Scrambled	769	103	97.4	4.7
Literal	735	123	94.8	4.9
	Low-IQ Group			
Metaphorical	834	117	94.7	5.4
Scrambled	823	121	97.0	4.2
Literal	751	129	92.0	6.1

Paired *t* tests revealed that the high-IQ group showed statistically reliable differences between the metaphorical and the scrambled sentences in both subject and item analyses [$t_s(15) = 2.27$, $p < .05$; $t_i(39) = 2.51$, $p < .05$], whereas the difference did not reach statistical significance for either the medium-IQ group [$t_s(15) = 1.68$, $p = .12$; $t_i(39) = 1.09$, $p = .28$] or the low-IQ group [$t_s(15) = 0.88$, $p = .39$; $t_i(39) = 0.40$, $p = .69$]. This result provides good evidence that automaticity of metaphor activation is not an all-or-none phenomenon. Rather, the higher the IQ of the reader, the more likely that metaphorical meaning is activated, which in turn creates interference in judging the metaphor to be literally untrue. Interestingly, as can be seen by the standard deviations reported in Table 2, the variability within groups was very similar. This suggests that the lack of significant differences for the low-IQ group was not due to increased variability as IQ decreased. Rather, the mean MIE decreased linearly across the groups.

ERP stimulus-locked analysis. Our next analysis focused on the stimulus-locked ERP data, which allowed us to explore the time course of the MIE and also to examine whether the individual differences seen in the behavioral data were reflected in the ERPs. Figure 1 shows the grand average ERPs for all 48 subjects to the three sentence conditions at midline. The ERPs elicited by the scrambled sentences clearly were more negative than those to literal sentences beginning at about 250 msec and lasting until about 600 msec, with the responses to the metaphors falling in between. The three conditions appear to overlap to a large extent from 600 to 1,000 msec. Beginning at about 1,000 msec, the ERPs to the metaphors show a positive enhancement that extends to the end of the recording epoch and is more visible at Pz than at Fz.

The critical difference between the metaphorical and the scrambled conditions was apparent in the earliest recording region, 200–400 msec. However, this effect varied among the groups, as can be seen in Figure 2. The ERPs to metaphors were more positive than those to the scrambled sentences for the high- (Figure 3) and medium-IQ groups. This was particularly clear at the posterior sites. In contrast, the low-IQ group (Figure 4) showed little difference between the metaphorical and the scrambled sentence conditions in the regions associated with the N400 component and, in contrast to the high- and medium-IQ groups, showed greater positivity to the scrambled than to the metaphorical condition for the later positivity (Figure 3).

To confirm these observations, a series of statistical analyses was computed to compare the sentence conditions, with both subjects and items as random factors. Although item analysis is a standard procedure within psycholinguistics, ERP researchers working with natural language stimuli have typically conducted only subject analyses. However, this limits the potential generalizability of the results to other linguistic utterances of the same type (Clark, 1973). The use of this procedure is particularly useful when the pool of linguistic stimuli is limited, and so we conducted an item analysis on the ERPs.[5] To quantify changes over the time course of processing, measurements were made for mean amplitude in 200-msec latency bins beginning at 200 msec. A series of ANOVAs was conducted to compare the sentence conditions, with Greenhouse–Geisser correction factors being used as necessary. Analyses were conducted first at midline sites; hemispheric comparisons were then conducted in separate analyses. For midline, a $2 \times 3 \times 3$ (sentence type $\times$ front/back electrode site $\times$ IQ group) mixed ANOVA was computed separately for each bin. Sentence type (metaphorical or scrambled) and front/back (Fz, Cz, or Pz) were repeated measures, and IQ group (high, medium, or low) was a between-subjects variable. To better characterize the differences across the scalp between ERPs to scrambled and metaphorical sentences, a $2 \times 5 \times 2 \times 3$ mixed ANOVA was computed for each latency bin, where sentence type (metaphorical or scrambled), front/back (frontal, central, parietal, temporal, or occipital), and laterality (left or right) were re-

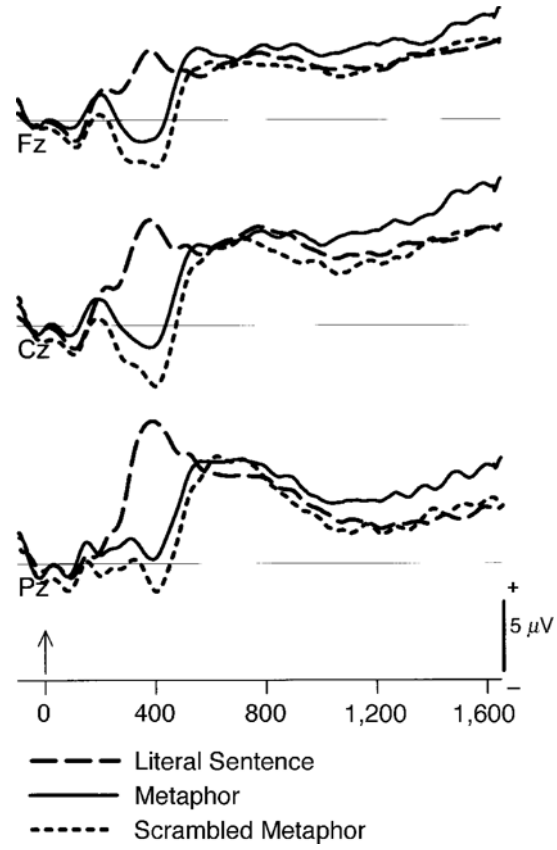

Figure 1. Grand average (n = 48) event-related potentials recorded from the three midline sites, elicited by the final word of literal, metaphorical, and scrambled sentences. Positive is plotted up in this and subsequent figures. The top of the figure corresponds to the front-central site, indicated by Fz. Cz refers to the mid-central site, and Pz refers to the parietal-central site. Onset of the final word of the sentence is indicated as 0 on the time line.

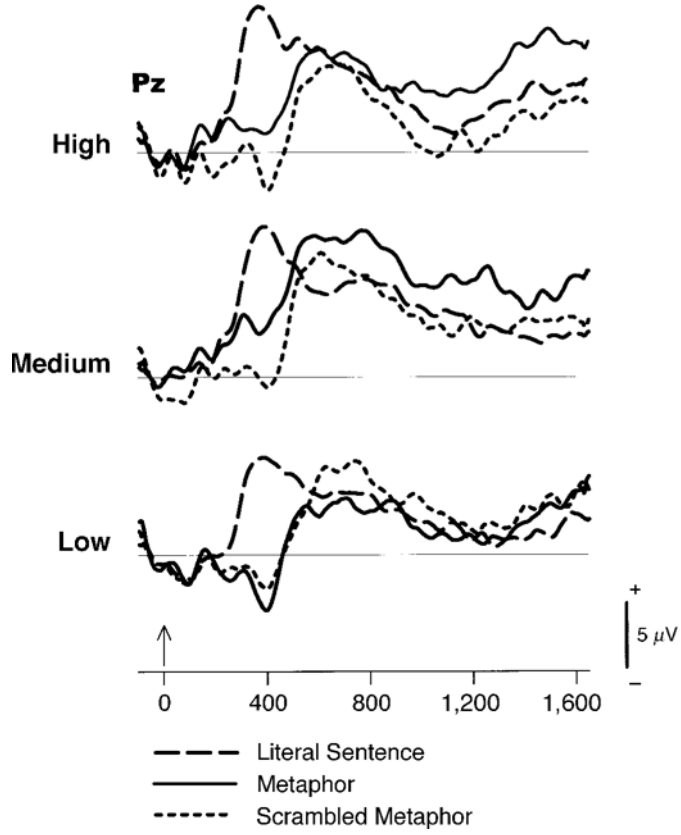

Figure 2. Grand average waveforms elicited by the final word of the literal (light dashed line), metaphorical (solid line), and scrambled (bold dashed line) sentences at the Pz electrode site for each IQ group.

peated measures and IQ group (high, medium, or low) was a between-subjects measure. The critical effects were those that focused on differences in sentence type (metaphorical or scrambled), and higher order interactions with sentence type were evident at most bins. These interactions were then explored with a series of planned comparisons, using paired *t* tests to compare metaphorical with scrambled conditions (the comparison that reflected the MIE) for each IQ group. All the differences described were significant ($p < .05$, two-tailed) for subject and item analyses, unless otherwise specified.

Early effects: 200–400 msec. ERPs were more negative to scrambled than to metaphorical sentences in this latency window. This main effect of sentence type was reliable both at midline [$F_s(1,45) = 6.35$, $MS_e = 36.89$, $p = .015$; $F_i(1,114) = 4.11$, $MS_e = 96.41$, $p = .045$] and in the laterality analysis [$F_s(1,45) = 6.54$, $MS_e = 101.36$, $p = .01$; $F_i(1,114) = 5.18$, $MS_e = 253.89$, $p = .025$]. At midline, the three-way interaction among sentence type, front/back, and group was significant for both subject and item analyses [$F_s(4,90) = 3.16$, $MS_e = 1.44$, $p = .025$; $F_i(4,228) = 2.67$, $MS_e = 1.72$, $p = .039$].

Planned comparisons showed that this interaction was a result of a significant difference in the subject analysis between the metaphorical and the scrambled conditions that was reliable for the high-IQ group at central and posterior sites (C3, C4, Pz, P3, P4, O1, O2, T5, and T6). For the medium-IQ group, the metaphorical–scrambled difference was reliable at all frontal, central, and parietal sites (for the subject analysis, $p < .05$ at all sites; for the item analysis, only at Pz). For the low-IQ group, the difference between the metaphorical and the scrambled conditions was not significant at any electrode site for either the subject or the item analysis.

Early effects: 400–600 msec. Similar to the results for 200–400 msec, mean amplitude for ERPs to scrambled stimuli was more negative than that for ERPs to metaphors, as can be seen in the main effect for sentence type at midline [$F_s(1,45) = 4.26$, $MS_e = 61.45$, $p = .045$; $F_i(1,114) = 3.13$, $MS_e = 154.75$, $p = .08$]. This was also true in the laterality analysis [$F_s(1,45) = 3.91$, $MS_e = 171.94$, $p = .05$; $F_i(1,114) = 3.15$, $MS_e = 421.06$, $p = .08$]. The main effect of group was significant in both midline [$F_s(2,45) = 3.69$, $MS_e = 142.92$, $p = .03$; $F_i(2,114) = 7.22$, $MS_e = 156.26$, $p = .001$] and laterality [$F_s(2,45) = 3.73$, $MS_e = 397.04$, $p = .03$; $F_i(2,114) = 6.33$, $MS_e = 477.28$, $p = .002$] analyses. These main effects were moderated by the three-way front/back × sentence type × group

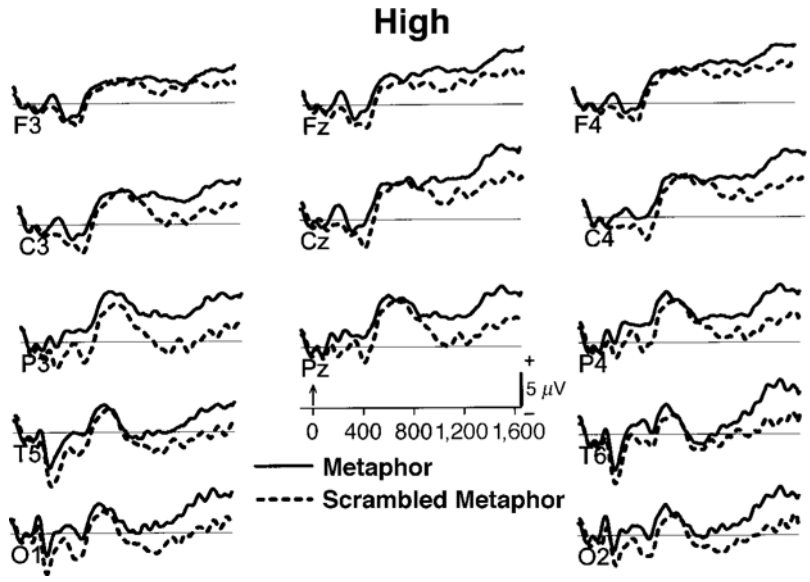

Figure 3. Grand average waveforms for the high-IQ group (n = 16) at 13 electrode sites, elicited by the final word of the metaphorical and scrambled sentences. The left column represents recordings from the left hemisphere, the center column from the midline, and the right column from the right hemisphere. Frontal sites are indicated by "F," central sites by "C," parietal sites by "P," temporal sites by "T," and occipital sites by "O."

interaction in both the midline [$F_s(4,90) = 2.77$, $MS_e = 2.41$, $p = .04$; $F_i(4,228) = 3.13$, $MS_e = 7.51$, $p = .02$] and the laterality [$F_s(8,180) = 2.58$, $MS_e = 11.80$, $p = .04$; $F_i(8,456) = 2.26$, $MS_e = 20.40$, $p = .08$] analyses.

Planned comparisons were performed as for the earlier latency bin. For the high-IQ group, ERPs to the scrambled sentences were more negative than those to the metaphors. These differences approached significance

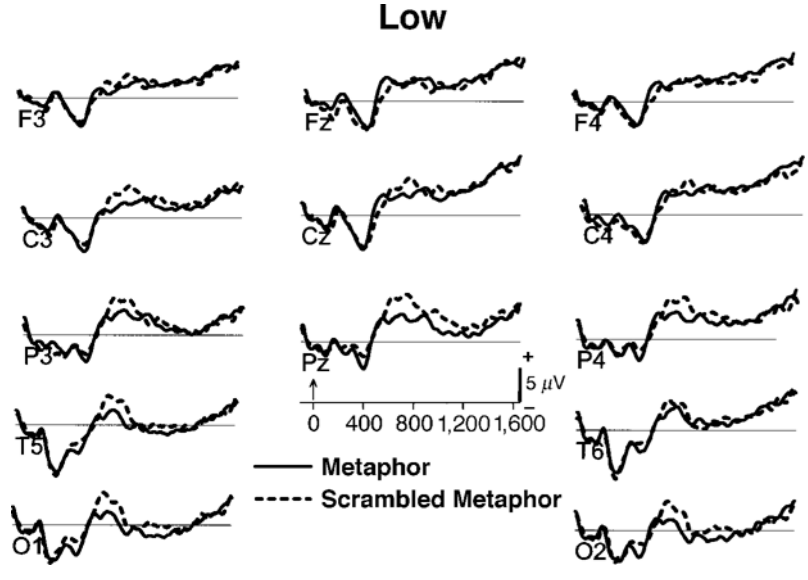

Figure 4. Grand average waveforms for the low-IQ group (n = 16) at 13 electrode sites, elicited by the final word of the metaphorical and scrambled sentences. Frontal sites are indicated by "F," central sites by "C," parietal sites by "P," temporal sites by "T," and occipital sites by "O."

($p < .10$) for the high-IQ group at P3, O1, and O2. The medium- and low-IQ groups showed no reliable differences between sentence types at any electrode site.

Middle effects: 600–800, 800–1,000, and 1,000–1,200 msec. In these middle latency bins that comprise the late positivity, the only reliable effect of sentence type was a two-way interaction of sentence type and laterality for the 800- to 1,000-msec bin in the subject analysis [$F_s(1,45) = 4.68, MS_e = 2.91, p = .036$; $F_i(1,114) = 2.99, MS_e = 12.33, p = .087$]. Overall, the metaphors showed more equal activation in both hemispheres (left = 3.2 μV, right = 3.8 μV) relative to the scrambled (left = 1.5 μV, right = 2.6 μV). This means that the left–right difference was greater for scrambled sentences than for metaphorical sentences. No other effects of sentence type or group were reliable in these latency windows. The planned comparisons revealed that the metaphorical–scrambled difference was reliable only for the high-IQ group (1,000–1,200 msec at Pz, P3, and O1; it approached significance at C3).

Late effects: 1,200–1,400 and 1,400–1,600 msec. In the midline analysis for 1,200–1,400 msec, the three-way sentence type × front/back × IQ group interaction was reliable for the subject analysis [$F_s(4,90) = 2.84, MS_e = 3.97, p = .001$], but not for the item analysis [$F_i(2,228) = 1.28, MS_e = 12.81, p = .283$]. A four-way sentence type × front/back × laterality × IQ group interaction was reliable in the laterality–subject analysis [$F_s(8,180) = 2.28, MS_e = 1.01, p = .03$], but not in the item analysis [$F_i(8,456) = 1.29, MS_e = 5.03, p = .27$]. For the 1,400–1,600 msec bin, in contrast to the previous bin, the three-way interaction was not reliable at midline. However, similar to the 1,200–1,400 msec bin, the four-way sentence type × laterality × front/back × IQ group interaction approached significance in the subject analysis [$F_s(8,180) = 2.06, MS_e = 1.28, p = .06$; $F_i(8,456) = 1.28, MS_e = 283.39, p = .27$]. Planned comparisons showed that the high-IQ group showed a significant difference between metaphorical and scrambled conditions at primarily posterior sites (1,200–1,400 msec, Pz, P3, with a trend evident at O1 and T6; 1,400–1,600 msec at C4, Pz, P3, P4, O1, O2, T5, and T6). Neither the medium-IQ nor the low-IQ group showed a reliable difference between sentence types at any electrode site in either subject or item analysis.

In summary, the appearance of an MIE, indexed by a distinct difference between ERP amplitudes in the metaphorical and the scrambled conditions, began around 200 msec, the region associated with the emergence of the N400 component. This difference was reliable for the medium- and high-IQ groups from 200 to 400 msec but, by 400 msec, was reliable only for the high-IQ group. There were no differences between the metaphorical and the scrambled conditions in the middle latency bins (600–1,200 msec, roughly comprising the late positive component); however, differences reemerged in the later bins (1,200–1,600 msec). The waveforms in Figure 1 show greater positivity of the metaphorical condition with respect to to the scrambled condition that is larger in posterior regions. This pattern was most clearly evident in the high-IQ group. In the 1,400–1,600 msec bin, the subject analysis showed a broad difference that encompassed both posterior regions in both hemispheres.

ERP response-locked analysis. As was shown in the behavioral analysis, the three groups differed in their overall RTs, with high-IQ subjects tending to be faster; therefore, the ERPs were reaveraged locked to RT. The response-locked ERPs align the waveforms on the basis of each subject's response and so should be a more sensitive measure of whether differences between the two sentence conditions continue to exist after response (Coles et al., 1995). Measurements were made of average amplitude in 100-msec bins for the 400 msec prior to response (indicated with a negative latency) and for the 600 msec after response. Separate ANOVAs were conducted for each latency bin, as was done for the stimulus-based analyses. The F, df, MS_e, and p values of the significant effects of or interactions with IQ group or sentence type are reported in Table 3. Results of t tests ($p < .05$, two-tailed) that compared the pairs of sentence types were conducted at each electrode site separately for each group. Maps displaying the sites of significant metaphorical–scrambled differences are shown in Figure 5 for the comparisons.

Preresponse bins: −400 to −300 and −300 to −200 msec. The ERPs in the two earliest bins showed similar results. The main effect of sentence type was significant at midline and in the laterality analyses. Post hoc tests showed that all pairwise comparisons were reliable, with amplitude being greatest for the literal condition and most negative for the scrambled condition. The sentence type × front/back interaction was significant for the midline and laterality analyses. The IQ group × sentence type interaction was significant for the −300- to −200-msec bin. These effects were moderated by the significant sentence type × laterality × IQ group interaction in the laterality analyses for both bins.

Planned comparisons using paired t tests were conducted separately for each IQ group. The ERPs to scrambled sentences showed the most negative waveforms. The analyses conducted separately for the −400- to −300-msec and −300- to −200-msec bins showed similar result patterns. The sentence conditions differed for nearly all the sites for the high-IQ group. The low-IQ group showed no reliable differences for the metaphorical–scrambled comparison. The difference approached significance at Cz, F3, and T5 in the −400- to −300-msec bin for the medium group.

−200 to 0 msec. The main effect of sentence type continued to be significant from −200 to −100 msec, but only approached significance for the 100 msec just prior to response. The distribution of the ERPs across the scalp varied between the sentence types for the two bins (sentence type × front/back and sentence type × laterality). The group × sentence type interaction approached significance for the −200- to −100-msec bin but disappeared just prior to response. The planned comparisons showed that the high-IQ group continued to show more

Table 3
Results of Analyses of Variance for the Response-Based ERP Analyses for the Sentence Type Main Effect and Significant Interactions With Sentence Type

Latency Bin	Analysis	Effect	*F*	*df*	Sig.	MS_e
−300 to −400	midline	sentence type	4.32	1,45	.043	42.26
	lateral	sentence type	8.01	1,45	.007	123.55
		sentence type × front/back	5.43	4,180	.008	6.99
		sentence type × lateral × IQ group	4.42	2,45	.018	3.93
		sentence type × front/back × lateral	2.91	4,180	.039	0.94
−200 to −300	midline	sentence type	2.93	1,45	.094	53.82
	lateral	sentence type	2.87	1,45	.097	169.31
−200 to −100	midline	sentence type	2.96	1,45	.092	61.99
	lateral	sentence type	2.45	1,45	.125	188.50
		sentence type × lateral × IQ group	2.34	2,45	.108	3.00
−100 to 0	midline	sentence type	2.48	1,45	.122	61.44
	lateral	sentence type	1.44	1,45	.236	198.36
0 to 100	midline	sentence type	0.26	1,45	.613	85.20
	lateral	sentence type	0.25	1,45	.617	247.01
100 to 200	midline	sentence type	0.09	1,45	.769	117.83
		sentence type × midline × IQ group	1.85	4,90	.151	5.20
	lateral	sentence type	0.03	1,45	.872	332.51
		sentence type × front/back × IQ group	2.59	8,180	.011	10.18
200 to 300	midline	sentence type	0.41	1,45	.525	110.44
	lateral	sentence type	0.43	1,45	.518	301.30
300 to 400	midline	sentence type	0.77	1,45	.386	106.95
	lateral	sentence type	0.99	1,45	.326	301.42

Note—*F*, significance (Sig.), and MS_e values are reported with Greenhouse–Geisser correction factors where appropriate. The *df* values are reported uncorrected.

positive ERPs to metaphorical than to scrambled sentences prior to response but that this difference was reliable only at F4 in the −200- to −100-msec bin.

Postresponse differences. The main effect of sentence type was not reliable for any of the postresponse analyses. The sentence type × front/back × IQ group interaction approached significance in the 100 msec immediately following response and was significant at midline for 100–200 msec and, in the laterality analysis, for 100–200 msec and 200–300 msec (see Table 3).

The planned comparisons revealed a reemergence of a metaphorical–scrambled difference, in the high-IQ group, as a greater positivity to metaphorical than to scrambled sentences at C4 and T6 in the 300- to 400-msec bin. The difference between the metaphorical and the scrambled conditions was not reliable at any postresponse bin for either the medium- or low-IQ group.

The response-locked analyses revealed that the MIE varied for the IQ groups and that this effect varied across time. The high-IQ group showed strong bilateral differences between the metaphorical and the scrambled conditions across the scalp from −400 to −300 msec preresponse, then no differences at the point of response, and a reemergence of the MIE that was strongest in the right hemisphere from 300 to 400 msec. In contrast, the low-IQ group showed no reliable differences between the metaphorical and the scrambled conditions at any of the bins at any of the electrode sites, with the medium-IQ group showing a trend toward a difference prior to response.

Regression Analysis

Consistent with other research (Holcomb & Neville, 1990), N400 amplitude tended to be largest toward the back of the head over medial sites. Therefore, in order to get a single measure of the MIE, the mean amplitudes for Pz, P3, and P4 were averaged separately for the metaphorical and the scrambled conditions in the bin that included the onset of N400 (200–400 msec in the stimulus-based analysis). The average score for the scrambled condition was then subtracted from the average score for the metaphorical condition to provide a measure of the MIE. Correlations with the individual difference variables showed that the MIE correlated with overall IQ (r = .34, p = .02) and was slightly higher with KBIT Verbal IQ (r = .39, p = .006), but not with the KBIT Matrices score (r = .22, p = .14). There was also no significant correlation between the Reading subtest of the Wide Range Achievement Test (WRAT; r = .18, p = .26) or the Word Attack subtest of the Woodcock Reading Mastery Test–Revised (r = .20, p = .21), where subjects had been asked to decode phonetically legal nonsense words. The reading tests were correlated with each other (r = .66, p < .001). A linear regression was conducted to investigate whether IQ could predict the ERP MIE. This revealed that the subjects' scores on the IQ test were a significant predictor of the size of their ERP MIE [$F(1,46) = 5.90$, p = .01]. IQ accounted for approximately 11% of the variance. The verbal IQ score explained approximately 15% of the variance [$F(1,46) = 8.26$, p < .01]. None of the other individual difference variables, such as age, sex, handedness, WRAT, or the Word Attack test of the Woodcock, added significant variance to the equation.

To summarize the key ERP findings, when we compare the two sets of analyses, we see very similar patterns. The MIE emerged very early in the waveform, disappeared in the middle bins, and then reemerged in the later bins. Even

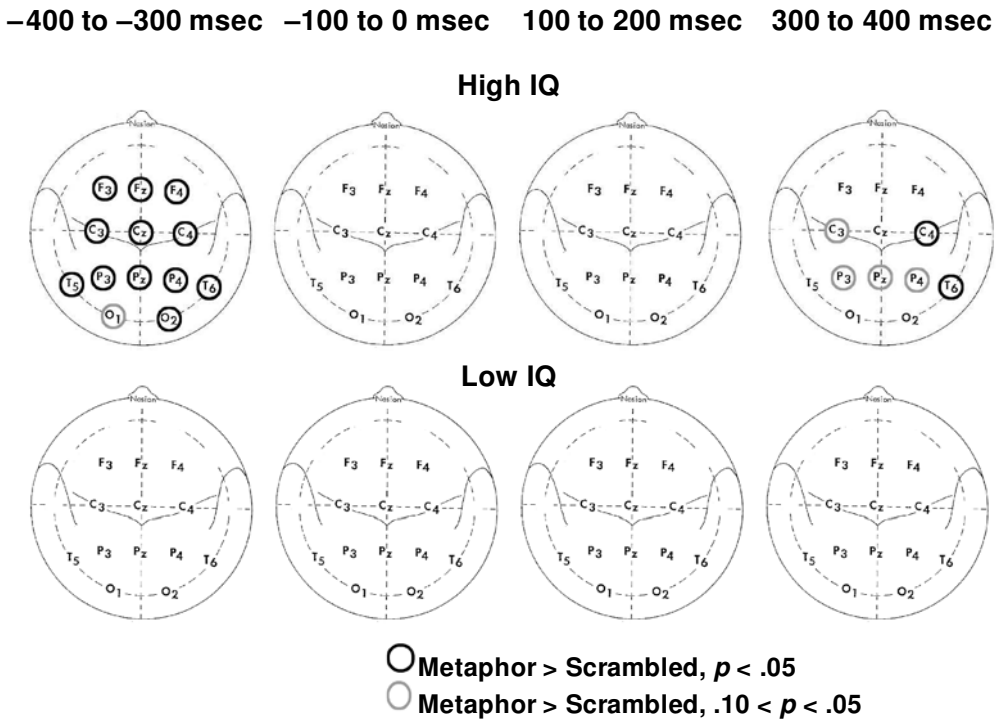

Figure 5. Planned comparison *t* tests of the response-related event-related potential (ERP) averages to metaphorical and scrambled sentences for the high- and low-IQ groups. Electrode sites at which the planned comparison showed greater (more positive) amplitude to metaphorical than to scrambled ERPs are indicated with a circle. Black circles indicate a difference at the $p < .05$ level; gray circles indicate a difference at the $.10 < p < .05$ level.

when aligned to each individual's behavioral response, the differences between the metaphorical and the scrambled conditions remained robust for the high-IQ group but unreliable for the low-IQ group. It is interesting that the late effects continued to be reliable for the high-IQ group even when aligned for RT. These late effects might be tentatively related to working memory operations at sentence wrap-up that are more effortful for low- than for high-IQ readers (Kutas, Federmeier, & Sereno, 1999).

EXPERIMENT 2

Experiment 1 successfully replicated the basic MIE and then showed that the effect was stronger in subjects with higher IQs. There are several possible explanations for why the understanding of metaphors is less automatic in those with lower IQs. It is possible that the low-IQ subjects are less verbal, read less, and so have had less exposure to these particular metaphors. If so, we might see a difference when we ask high- and low-IQ subjects to rate the metaphors for familiarity. Low-IQ subjects might also have poorer basic vocabularies and, perhaps, less rich semantic networks. Most IQ tests, including those used here, have a vocabulary component.

It is also possible that those with lower IQs have capacity limitations that impair comprehension. Working memory has been shown to correlate with measures of general intelligence (Engle, Kane, & Tuholski, 1999). Our lower IQ subjects might have required more of their available resources for the word-by-word reading task, which, in turn, limited the resources available for linking the more peripheral semantic features of topic and vehicle often required for understanding of metaphors. If this is true, lower IQ subjects might be more successful at the comprehension of metaphors in an untimed task.

In Experiment 2, we examined these possibilities in an off-line judgment and rating task. We presented the same set of 40 metaphors and asked the subjects to rate them for familiarity and ease of comprehension, as well as to supply a written interpretation. We also collected a set of individual difference variables, including spatial and language working memory span and separate vocabulary and comprehension scores, to begin to identify some of the component processes that might account for the individual differences seen in Experiment 1.

Method

Subjects. Thirty-four undergraduate college students participated in the study (60% female). All were native speakers of English and reported no reading disabilities. They received either course credit or a $5 payment for their participation.

Materials and Stimuli. The stimuli were the same set of 40 metaphors as that used in Experiment 1. The metaphors were pre-

sented in written form in the same random order for each subject. The subjects read each metaphor and rated (1) how easy it was to comprehend (1 = *low* to 7 = *high*), and (2) how familiar is was (1 = *not familiar* to 7 = *very familiar*). Finally, they were asked to write out their interpretation of each metaphor. They were also told that if they did not understand a metaphor, they could leave it blank.

The subjects completed the verbal IQ section of the Multidimensional Aptitude Battery, Version II (MAB; Jackson, 1998). The MAB was used instead of the KBIT for two reasons: The MAB can be administered to groups, and it allows the calculation of separate vocabulary and comprehension scores. Although both IQ tests correlate highly with the WAIS, we wanted to be sure that the scores from Experiments 1 and 2 were comparable; therefore, 18 of the 34 subjects in Experiment 2 were also individually administered the KBIT. The correlations between KBIT scores and MAB verbal IQ scores were high ($r = .69$, $p < .001$). Thus, we grouped our subjects on the basis of the same criteria as those used in Experiment 1: 115 and above, high IQ; 114 to 100, medium IQ; and 99 and below, low IQ. In the analyses reported below we will report the MAB comprehension scale and the MAB vocabulary scale separately.

All the subjects also completed measures of verbal and spatial working memory. Some research (e.g., Shah & Miyake, 1996) suggests that these two forms of working memory may rely on separate pools of resources. Therefore, we would expect working memory for language to be a better predictor of metaphor comprehension than spatial working memory would be. By including the spatial working memory task, we could look for problems of common method variance, in the event that all the measures were highly intercorrelated. Verbal working memory was assessed using a listening span task developed by Tompkins, Bloise, Timko, and Baumgaertner (1994) and based on Daneman and Carpenter (1980). The subjects listened to a series of sentences and judged whether they were true or false. At the end of each trial, they were asked to recall the last word of all the sentences in the set. The set size increased from two to five sentences per set. Each set had three trials. Working memory span was scored as the largest set size for which the listener remembered all of the last words in at least two of the three trials for the set. A half point was added for one of the three. For example, if a subject remembered all of the words in all three trials at set size 3 and, then, one of the trials in set size 4, their working memory span would be 3.5. The task was presented on a Macintosh computer using the PsyScope experimental program (Cohen, MacWhinney, Flatt, & Provost, 1993). The spatial working memory test, developed by Shah and Miyake (1996), has similar logic. Subjects judge whether a rotated letter is normal or backward. At the end of the set, the subjects must recall the locations of the top of each letter by pointing to a grid. The set size increases from two to five letters per trial. There are five trials per set, and the spatial span is scored as the highest set for which at least three trials had all of the letter locations recalled correctly. The task was programmed using E-prime (Schneider, Eshman, & Zuccolotto, 2002) and was run on a 500-mHz computer with Win98.

Procedure. All the subjects provided informed consent. They were then run in small groups of 5 to 10 individuals. They first completed the two working memory tasks and the MAB verbal test. They then rated and interpreted the metaphors.The whole experiment took about 60 min. Eighteen of the 34 subjects also were individually administered the KBIT IQ test in a separate 40-min session.

Scoring. The metaphor interpretations were judged for quality by two trained undergraduate research assistants who were unaware of the hypotheses and blind to the writers' identities and scores. Each rater was trained on the scoring criterion by the authors until they reached high levels of interrater reliability (percentage of agreement, 83%; $r = .79$, $p < .001$). The quality scale that was used ranged from 1 (*no interpretation* or *very poor quality interpretation*) to 5 (*very high quality interpretation*). Each rater was trained on the average interpretatio ns of each metaphor (scored as 3). Higher scores were given for more in-depth or multiple interpretations. Take, for example, the metaphor, *the puppies were a tornado*. The interpretation "they destroyed everything" was given a score of 3, but the interpretation "they rushed about in circles and made a huge mess of the place" was scored a 4, because more features of the metaphor were discussed. The quality ratings of the two judges were then averaged for all the analyses.

Results

Pearson *r* correlations were used to examine the relationships between the individual difference variables, metaphor ratings and metaphor interpretations. One important question was whether the IQ measure used in Experiment 1 would correlate with working memory for language. The results of the 18 subjects who completed both the KBIT and the working memory measures confirmed that there was a positive relationship between the measures ($r = .69$, $p < .001$).

One-way ANOVAs were conducted to assess differences between the IQ groups. There were no statistically significant between-group differences on ratings of metaphor comprehensibility [$F(2,30) = 2.17$, $p = .13$] or subjective familiarity [$F(2,31) = 2.92$, $p = .07$]. However, the IQ groups differed on the rated quality of their interpretations [$F(2,31) = 4.01$, $p = .02$]. In addition, the high-IQ group had significantly fewer very poor quality or missing interpretations (3.5%) than did the medium- (8.6%) or the low-IQ (9.5%) groups.

Table 4 shows the correlations between the ratings and the individual difference variables. Consistent with the previous analysis, the ratings of comprehension and familiarity did not correlate significantly with the individual difference measures. However, the independent ratings of interpretation quality were related to MAB vocabulary ($r = .58$, $p < .01$) and MAB comprehension ($r = .40$, $p < .05$). There was also a significant positive relationship between the verbal working memory measure and the MAB comprehension measure ($r = .41$, $p < .05$). As was expected, the measure of spatial working memory was not correlated with the verbal IQ measures or metaphor interpretations.

Discussion

The data clearly show that low-IQ subjects similar to those who participated in Experiment 1 were capable, given enough time, of understanding the metaphors. The low-IQ subjects did not differ on their self-reported ratings of familiarity or ease of comprehension, and they provided adequate interpretations for over 90% of the metaphors. However, they did differ on the quality of their interpretations. This fits well with previous research suggesting that task demands interact with subject characteristics to determine performance on metaphor comprehension (Blasko, 1999).

One possible explanation for the reduction of the MIE in the low-IQ subjects of Experiment 1 was that they might have smaller vocabularies and, therefore, potentially less rich semantic networks, leading to less automatic activation of the metaphor. There was some support for this

Table 4
Experiment 2: Pearson *r* Correlations Between Individual Difference and Rating Variables

	1	2	3	4	5	6
1. MAB vocabulary						
2. MAB comprehension	.48**					
3. Verbal working memory	.29	.41*				
4. Spatial working memory	.07	.11	.32			
5. Familiarity	−.19	−.13	−.13	−.12		
6. Comprehension ease	.27	.34	.00	−.06	.32	
7. Interpretation quality	.58**	.40*	.23	.03	.17	.32

*$p < .05$. **$p < .01$.

view. Although the vocabulary scale of the MAB was not significantly correlated with self-reported ratings of either familiarity or ease of comprehension, it did correlate with independent ratings of the quality of the interpretations.

Overall, the results suggest that both vocabulary and working memory capacity may have played a role in constraining comprehension in both experiments. They may have played a role in the lack of an MIE for the low-IQ group, if, as we suspect, the word-by-word reading task used more of the subjects' available capacity, whereas in Experiment 2, the rating and interpretation task provided more time to consider the meaning of the metaphor. The comprehension rating task required only that the reader be aware that the metaphor had some meaning, but the interpretation task also required them to explain exactly how it was meaningful.

Although the MAB vocabulary scale did not relate directly to the verbal working memory measure, the MAB comprehension scale was related to both quality of interpretation and verbal working memory. This suggests that vocabulary skill alone is not enough to explain the individual differences seen here. The comprehension test items asked the subjects to integrate the meanings of various sentences and to draw inferences about the intent of the author, processes that may require more working memory. Therefore, it is plausible that some of the individual differences seen in Experiment 1 may have been due to vocabulary differences, but comprehension and working memory capacity are likely to have played roles as well.

GENERAL DISCUSSION

There were three major goals of the present work. First, we wished to replicate the MIE and used ERP recordings to map its time course and pattern of activity across the scalp. Second, we wanted to determine whether the MIE was sensitive to individual differences in IQ. Finally, in Experiment 2, we used a rating and interpretation task to investigate some of the potential reasons for these individual differences.

Experiment 1 did replicate the MIE; overall, the metaphors were more difficult to reject as literally untrue than were the scrambled controls consisting of the same topic and vehicle terms. This result was reflected in the ERP analysis by a divergence in the averaged waveforms of the metaphorical and scrambled conditions such that the scrambled condition showed a larger N400 component. If the size of the N400 reflected response difficulty, we would have seen larger N400 amplitude for metaphorical than for scrambled sentences. But these data fit better with the interpretation of N400 as reflecting semantic integration. Consistent with this idea, it may be that the smaller N400 for the metaphorical sentences reflects an automatic semantic activation of the meaning of the metaphor that is not present for the scrambled items. The ERPs also showed that activation of the metaphor was almost immediately dampened so that, by the 400- to 600-msec bin, the effects were already statistically marginal. This could reflect the categorization of metaphors as "not literally true." Importantly, although the behavioral responses took approximately 800 msec, the ERPs showed that the metaphorical and scrambled waveforms began to diverge very early, only 200 msec after presentation of the last word of the sentence. This result is strongly supportive of a direct model of metaphor processing, in which figurative meaning is extracted automatically from interaction of the topic and the vehicle. However, much later in the time course, beginning around 1,200 msec, the waveforms for the metaphorical and the scrambled conditions began to diverge again in what we might call a *metaphor rebound effect.*

There are at least two possible explanations for this pattern. It is possible that the metaphorical meaning is inhibited in order to make the *no* response but that this inhibitory process is limited either by time or by necessity, since resources are required for other tasks. Once the decision has been made, inhibition might be released, and residual activation might reemerge. This seems to have at least anecdotal support. In several cases, the subjects mentioned that the meaning of a metaphor seemed to "pop" into consciousness after the decision was made. The second possibility is that the overall waveforms are the result of two different neurolinguistic systems working in parallel. One system leads to a quick activation and inhibition, and the second leads to a slower, more gradual activation of peripheral meaning (e.g., Beeman, 1993; Titone, 1998). These two possibilities are impossible to distinguish here but could be the focus of future work.

The second major goal was to investigate the possibility of individual differences in the MIE. Higher IQ

subjects were much more likely to show interference. For them, the metaphorical meaning appeared to be automatically activated, thereby making the decision to reject a metaphorically true sentence all the more difficult. Low-IQ subjects, on the other hand, had no trouble with the task; they acted as if both metaphors and scrambled sentences were equally untrue. This second finding seems more consistent with indirect models, in which figurative language processing is seen as a secondary and optional process that is more processing intensive than is literal language understanding.

There are several possible reasons for these differences. First, the results may be due, in part, to task demands. Although often used in ERP studies, the RSVP procedure used here may have made processing more difficult for low-IQ readers, because they had to remember each word long enough to make a decision. Lower IQ readers may simply have been using their more limited resources in the most efficient way possible. In their case, the metaphorical meaning was not highly activated, because they were using the available resources to accomplish the task at hand—reading the sentences and judging literal truth. It is possible that the high-IQ subjects had more available resources that allowed early activation of the appropriate features for metaphor interpretation, followed by efficient suppression of this unneeded information (Gernsbacher & Faust, 1991). Although metaphors and scrambled sentences clearly differed on their degree of meaningfulness, the task required that they be grouped into the *literally untrue, therefore, respond no* category. The response-locked ERPs seem to reflect this shift. For high-IQ individuals, there were clear differences across the entire scalp between both metaphorical and scrambled conditions of −400 to −300 msec. However, by the point of response, the metaphorical and scrambled sentences no longer differed; both were ready to be judged as literally untrue. The MIE then reemerged in the metaphor rebound effect 300–400 msec after the response.

Things are very different for the low-IQ readers. Just as beginning and very poor readers show little or no interference in a Stroop color–word task (e.g., Stanovich, Cunningham, & West, 1981), metaphor processing for low-IQ/low–working-memory readers may be optional in cases in which resources are needed elsewhere. This does not mean that these readers are incapable of understanding the metaphors used here. As was shown in Experiment 2, over 90% of the same metaphors as those used in Experiment 1 were interpreted by low-IQ subjects in an off-line task, and the IQ groups did not differ in their self-reported ratings of comprehensibility. But, when presented word-by-word and when strategically focused away from the processing of metaphorical meaning, the low-IQ group showed little or no activation. An important issue for future research is whether this lack of activation for the lower IQ group was the result of capacity constraints alone or whether the underlying semantic representations that are called upon in the processing of metaphors also differ. In the latter case, the central meanings of the topic and vehicle terms themselves may be activated, but the conceptual representations may be relatively impoverished. In contrast, the representations may be similar, but the degree of spreading activation needed to uncover the peripheral characteristics of the terms that are appropriate to the figurative meaning may be sensitive to resource limitations. These two possibilities are not mutually exclusive. We know from the results of Experiment 2 that even when given unlimited time to deliberately interpret the metaphors, the low-IQ group produced poorer quality interpretations.

Neither direct nor indirect models of metaphor comprehension can completely explain the findings. Therefore, we propose abandoning this distinction and, instead, developing a new approach based on the concept of constraint satisfaction. Constraint satisfaction is the assumption that, in natural language understanding, all of the information in the discourse and the social setting (including the demands of the task) will provide constraints on understanding. These different sources of information may conflict, and if so, the alternative meanings will compete for activation over time, with those constraints providing probabilistic evidence in support of various alternative interpretations. Competition ends when one alternative is the best fit (Katz & Ferretti, 2001). We would like to add a seemingly obvious point, that comprehension also depends very much on the particular mental representations of the comprehender that have been built over time by social/cultural experiences. In addition, processing a particular utterance, whether literal or nonliteral, requires varying degrees of mental effort, as implemented in the underlying neural resources. Consistent with capacity theory, individuals differ in the amount and efficiency of working memory capacity, so comprehension of a particular sentence may be constrained if the capacity limitations induced by the situation exceed the resources of the individual. This, of course, provides a much more complex, if more realistic, account of comprehension.

Theoretical efforts in this direction have already been made. For example, Katz and Ferretti (2001) have taken a constraint satisfaction approach to the understanding of how literal and figurative context influences the interpretation of proverbs. Similarly, Kintsch (1998, 2000, 2001) has taken a basic constraint satisfaction approach in his recent model of language comprehension, the predication model. The predication model may provide an explanatory framework for our findings, so it is worth explaining its basic characteristics. The model has two basic components, with the first being a model of human knowledge representation based on latent semantic analysis (LSA). In LSA, the co-occurrences of words in context, taken from large quantities of written text corpora, are automatically constructed into a high-dimensional semantic space. The structure of the semantic space is constantly changing as a function of the input. This model nicely captures semantic relationships, but not the

asymmetry of metaphors—for example, *the surgeon is a butcher* has a different meaning from *the butcher is a surgeon.* The second part of the theory, the construction integration (CI) model (Kintsch, 1998), accounts for this asymmetry. The CI model suggests that meaning activation spreads in time across a self-inhibitory conceptual network. For example, when one encounters the metaphor, *the lawyers are sharks*, the semantic neighborhoods of *lawyers* and *sharks* are activated. The semantic neighborhood of the topic constrains (or predicates) which semantic features of the vehicle (i.e., the predicate) will be relevant for interpretation of the metaphor. Concepts strongly related to *sharks* (e.g., fins) that cannot be integrated with those of *lawyers* will be inhibited by the semantic neighbors of *sharks* that are relevant to *lawyers* (e.g., vicious).

Although not discussed as such, the model may provide the theoretical mechanism for two types of individual differences in metaphor understanding. The first is provided by LSA itself. The corpus consists of the written text that, typically, is read by a student who has completed high school. However, unlike a computer model, individuals will vary widely in the amount and type of reading they have done, not to mention their verbal conversations. Therefore, one would expect that people would vary widely in the richness and development of their semantic networks. If the conceptual representations of either topic or vehicle are missing or impoverished, comprehension will suffer. This may explain processing differences between metaphors that are very novel, but for the relatively familiar metaphors used here, this would be less likely, because every college student would be expected to have had some exposure to the concepts. However, there is also a processing parameter that might be involved. For every application of the predication model, the size of the semantic neighborhood (m) to be searched must be set. The set size for literal predication is usually the 50 nearest neighbors; however, for metaphor processing, it must be set much higher (usually $m = 500$), before relevant concepts are found. This is because understanding a metaphor often requires identifying concepts that are much more semantically distant. If the size of the semantic neighborhood that must be searched is 10 times larger, it is conceivable that more resources would be required. If these are analogous to working memory, individual differences in the size of the resource pool or the efficiency of the processes might constrain comprehension. To take this one step further, low working memory individuals may deal with processing constraints either by strategically setting a smaller search space or by simply running out of the available resources regardless of whether the search is successful. In either case, they might be less likely to show automatic activation if the task is demanding, as in Experiment 1, and may be less likely to complete an adequate interpretation in off-line tasks, as in Experiment 2. Admittedly, this is going well beyond the theory as stated, but it does seem to fit both the present data and a large body of other, seemingly conflicting findings in the literature.

Future work should examine whether the constraint satisfaction approach is indeed the hybrid model needed to explain conflicting results in the literature that sometimes support direct processing approaches and sometimes indirect approaches. If the predication model can be elaborated to handle both the influence of social/cultural influences and individual differences in language comprehension, it might provide the computational framework for significant theoretical and empirical advances. However, to provide a comprehensive framework, it will also need to be neurologically informed, so additional research into the neural mechanisms underlying figurative language is needed.

REFERENCES

Beeman, M. (1993). Semantic processing in the right hemisphere may contribute to drawing inferences from discourse. *Brain & Language*, **44**, 80-120.

Beeman, M., Friedman, R. B., Grafman, J., Perez, E., Diamond, S., & Beadle Lindsay, M. (1994). Summation priming and coarse semantic coding in the right hemisphere. *Journal of Cognitive Neuroscience*, **6**, 26-45.

Bihrle, A. M., Brownell, H. H., & Powelson, J. A. (1986). Comprehension of humorous and nonhumorous materials by left and right brain-damaged patients. *Brain & Cognition*, **5**, 399-411.

Blasko, D. G. (1999). Only the tip of the iceberg: Who understands what about metaphor? *Journal of Pragmatics*, **31**, 1675-1683.

Blasko, D. G., & Briihl, D. S. (1997). Reading and recall of metaphorical sentences: Effects of familiarity and context. *Metaphor & Symbol*, **12**, 261-285.

Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 295-308.

Bottini, G., Corcoran, R., Sterzi, R., Paulesu, E., Schenone, P., Scarpa, P., Frackowiak, R. S. J., & Frith, C. D. (1994). The role of the right hemisphere in the interpretation of figurative aspects of language: A positron emission tomography activation study. *Brain*, **117**, 1241-1253.

Burgess, C., & Chiarello, C. (1996). Neurocognitive mechanism underlying metaphor comprehension and other figurative language. *Metaphor & Symbolic Activity*, **11**, 67-84.

Clark, H. H. (1973). The language as fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior*, **12**, 335-359.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, **25**, 257-271.

Coles, M. G. H., Smid, H., Scheffers, M., & Otten, L. (1995). Mental chronometry and the study of human information processing. In M. D. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of mind: Event-related brain potentials and cognition* (pp. 86-131). New York: Oxford University Press.

Connine, C. M., Blasko, D. G., & Wang, J. (1994). Vertical similarity in spoken word recognition: Multiple lexical activation, individual differences, and the role of sentence context. *Perception & Psychophysics*, **56**, 624-636.

Coulson, S. (2001). *Semantic leaps*. New York: Cambridge University Press.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, **19**, 450-466.

Engle, R., Kane, M., & Tuholski, S. (1999). Individual differences in

working memory capacity and what they tell us about controlled attention, general fluid intelligence and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory* (pp. 102-134). Cambridge: Cambridge University Press.

Fabiani, M., Gratton, G., & Coles, M. (2000). Event-related brain potentials: Methods, theory and application. In J. Cacioppo, L. Tassinary, & G. Brentson (Eds.), *Handbook of psychophysiology* (2nd ed., pp. 53-84). New York: Cambridge University Press.

Gentner, D., & Wolff, P. (1997). Alignment in the processing of metaphor. *Journal of Memory & Language*, **37**, 331-355.

Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 245-262.

Gernsbacher, M. A., & Keysar, B. (1995, November). *The role of suppression in metaphor interpretation.* Paper presented at the 36th Annual Meeting of the Psychonomic Society, Los Angeles.

Gibbs, R. W., Jr. (1994). *The poetics of mind: Figurative thought, language, and understanding.* Cambridge: Cambridge University Press.

Gildea, P., & Glucksberg, S. (1983). On understanding metaphor: The role of context. *Journal of Verbal Learning & Verbal Behavior*, **22**, 577-590.

Glucksberg, S. (2001). *Understanding figurative language.* New York: Oxford University Press.

Glucksberg, S., Gildea, P., & Bookin, H. B. (1982). On understanding nonliteral speech: Can people ignore metaphors? *Journal of Verbal Learning & Verbal Behavior*, **21**, 85-98.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, **97**, 3-18.

Gratton, G. (1997). *PC-EXP* [Computer software]. Columbia: University of Missouri.

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography & Clinical Neurophysiology*, **55**, 468-484.

Gregory, M., & Mergler, N. (1990). Metaphor comprehension: In search of literal truth, possible sense, and metaphoricity. *Metaphor & Symbolic Activity*, **5**, 151-173.

Hirst, W., LeDoux, J., & Stein, S. (1984). Constraints on the processing of indirect speech acts: Evidence from aphasiology. *Brain & Language*, **23**, 26-33.

Holcomb, P. J., & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language & Cognitive Processes*, **5**, 281-312.

Honeck, R. P., Welge, J., & Temple, J. G. (1998). The symmetry control in tests of the standard pragmatic models: The case of proverb comprehension. *Metaphor & Symbol*, **134**, 257-273.

Inhoff, A. W., Lima, S. D., & Carroll, P. J. (1984). Contextual effects on metaphor comprehension in reading. *Memory & Cognition*, **12**, 558-567.

Jackson, D. N. (1998). *Multidimensional aptitude battery II*. London, ON: Research Psychologist Press.

Jastak, S., & Wilkinson, G. S. (1984). *The Wide Range Achievement Test–Revised*. Wilmington, DE: Jastak Associates.

Johnson, J., & Pascual-Leone, J. (1989). Developmental levels of processing in metaphor interpretation. *Journal of Experimental Child Psychology*, **48**, 1-31.

Johnson, R. (1993). On the neural generators of the P300 component of the event-related potential. *Psychophysiology*, **30**, 90-97.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, **99**, 122-149.

Katz, A., & Ferretti, T. (2001). Moment-by-moment reading of proverbs in literal and nonliteral contexts. *Metaphor & Symbol*, **16**, 193-221.

Kaufman, A. S., & Kaufman, N. J. (1990). *Kaufman Brief Intelligence Test*. Los Angeles: Western Psychological Services.

Kazmerski, V., & Friedman, D. (1998). The scalp topography of P3b in early Alzheimer's disease. *Journal of Psychophysiology*, **12**, 127-143.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory & Language*, **30**, 580-602.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* New York: Cambridge University Press.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, **7**, 257-266.

Kintsch, W. (2001). Predication. *Cognitive Science*, **25**, 173-202.

Kutas, M. (1997). Views on how the electrical activity the brain generates reflects the functions of different language structures. *Psychophysiology*, **34**, 383-398.

Kutas, M., Federmeier, K. D., Coulson, S., King, J. W., & Münte, T. F. (2000). Language. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (2nd ed., pp. 576-601). New York: Cambridge University Press.

Kutas, M., Federmeier, K. D., & Sereno, M. I. (1999). Current approaches to mapping language in electromagnetic space. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 359-392). New York: Oxford University Press.

Light, L., Owens, S. A., Mahoney, P., & La Voie, D. (1993). Comprehension of metaphors by young and older adults. In J. Cerella, J. M. Rybash, W. J. Hoyer, & M. Commons (Eds.), *Adult information processing: Limits on loss* (pp. 459-488). San Diego: Academic Press.

Lovrich, D., Cheng, J. C., Velting, D. M., & Kazmerski, V. (1997). Auditory ERPs during rhyme and semantic processing: Effects of reading ability in college students. *Journal of Clinical & Experimental Neuropsychology*, **19**, 313-330.

Osterhout, L. (1994). Event-related brain potentials as tools for comprehending language comprehension. In C. Clifton, Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 15-44). Hillsdale, NJ: Erlbaum.

Osterhout, L., Bersick, M., & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, **25**, 273-285.

Osterhout, L., & Holcomb, P. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory & Language*, **31**, 785-806.

Pynte, J., Besson, M., Robichon, F., & Poli, J. (1996). The time course of metaphor comprehension. *Brain & Language*, **55**, 293-316.

Schneider, W., Eshman, A., & Zuccolotto, A. (2002). *E-Prime user's guide.* Pittsburgh: Psychology Software Tools.

Searle, J. (1979). Metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 83-111). Cambridge: Cambridge University Press.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, **125**, 4-27.

Stanovich, K. E., Cunningham, A. E., & West, R. F. (1981). A longitudinal study of the development of automatic recognition skills in first graders. *Journal of Reading Behavior*, **13**, 57-74.

Stelmack, R., & Houlihan, M. (1995). Event-related potentials, personality, and intelligence: Concepts, issues, and evidence. In D. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence: Perspectives on individual differences* (pp. 349-365). New York: Plenum.

Titone, D. (1998). Hemispheric differences in context sensitivity during lexical ambiguity resolution. *Brain & Language*, **65**, 361-394.

Tompkins, C. A., Bloise, C. G., Timko, M. L., & Baumgaertner, A. (1994). Working memory and inference revision in brain damaged and normally aging adults. *Journal of Speech & Hearing Research*, **37**, 896-912.

Trick, L., & Katz, A. N. (1986). The domain interaction approach to metaphor processing: Relating individual differences and metaphor characteristics. *Metaphor & Symbolic Activity*, **1**, 185-213.

Winner, E., & Gardner, H. (1977). The comprehension of metaphor in brain-damaged patients. *Brain*, **100**, 717-729.

Wolff, P., & Gentner, D. (2000). Evidence for role-neutral initial processing of metaphors. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 529-541.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests–Revised*. Circle Pines, MN: American Guidance Service.

NOTES

1. This example come from the beginning of the play, where Norfolk discloses the scheming of the Cardinal of York to Buckingham:

> Norfolk: All this was order'd by the good discretion
> Of the right reverend Cardinal of York.
>
> Buckingham: The devil speed him! no man's pie is freed
> From his ambitious finger. What had he
> To do in these fierce vanities? I wonder
> That such a keech can with his very bulk
> Take up the rays o' the beneficial sun
> And keep it from the earth.

2. Metaphors were chosen on the basis of an earlier normative study in which 20 different subjects rated 68 metaphors for aptness (metaphor goodness) and comprehension (ease of understanding), using a rating scale of 1 (*low*) to 7 (*high*). The stimuli averaged 5.07 for aptness and 5.21 for familiarity. In a second normative study, the metaphorical, scrambled, and literal sentences were judged for meaningfulness by 10 additional subjects on a scale of 1 (*low*) to 7 (*high*). The literal sentences ($M = 6.74$) and the metaphorical sentences ($M = 4.19$) were both judged as being more meaningful than were the scrambled sentences ($M = 1.60$).

3. Because N400 has been shown to be related to the predictability of the words in a sentence, we conducted a cloze probability study on the stimuli. Twenty-two subjects were presented with the beginning of each sentence and then were asked to complete the sentence. The literal sentences had a cloze probability of .56. Importantly, both the metaphorical and the scrambled sentences had essentially no predictability (metaphorical = .0056, scrambled = .0011). Therefore, any differences that might be found between these conditions was not due to the predictability of the stimuli.

4. The Reading subtest of the Wide Range Achievement Test–Revised (Jastak & Wilkinson, 1984) requires subjects to read single words aloud. This test takes about 5 min to complete and is administered individually. A grade equivalent score is obtained. In the Word Attack subtest of the Woodcock Reading Mastery Test–Revised (Woodcock, 1987), subjects decode phonetically legal nonsense words. These tests were used to ensure that none of the subjects had a significant reading disability.

5. The reasons given for the lack of item analyses are usually practical ones. The most serious is that in order to obtain an adequate signal-to-noise ratio, ERPs must be averaged over a large number of trials and, therefore, a larger number of subjects than is typically run in a condition would be needed (Osterhout, 1994). Yet many studies have been published with averages based on approximately 20 items, and the results have been found to be both replicable and stable. In the present work, we considered this an empirical issue and conducted item analyses in an exploratory manner on the stimulus-based analysis. If the item analyses lead to unstable averages, we should see consistently nonsignificant effects in the item analyses even in cases in which subject analyses are clearly significant. If the effects are significant in the item analyses, it provides an indication of stability across items. If both subject and item analyses are significant, it should provide confidence that the effect is stable across individuals and stimuli. Other major disadvantages to using item analyses in ERP language studies have included the need for tremendous amounts of disk space and computer memory (Osterhout, 1994). These are issues that are well known to behavioral and eye-tracking researchers but are now much less of a problem with the advent of newer, more powerful computers and higher capacity hard drives. It is true that item analysis is highly time consuming and, therefore, costly in and of itself, but we would like to disagree with the suggestion that it can be eliminated in favor of multiple replications across different sets of items (Osterhout, 1994). In fact, replication of an entire experiment with different stimuli is rarely found in the published literature and is even more time consuming than conducting an item analysis in the first place.

(Manuscript received October 6, 2000;
revision accepted for publication March 14, 2003.)