EUROPEAN JOURNAL OF COGNITIVE PSYCHOLOGY
2010, 22 (4), 529–568

Psychology Press
Taylor & Francis Group

Miscomprehension, meaning, and phonology: The unknown and phonological Armstrong illusions

Meredith A. Shafto and Donald G. MacKay
Psychology Department, University of California, Los Angeles, CA, USA

People often fail to detect the anomalous word in questions such as *How many animals of each kind did Moses take on the Ark?*, and incorrectly answer "two" despite knowing that Noah rather than Moses launched the Ark. The current study tests an account of this "Moses illusion" in which Moses mistakes reflect miscomprehension of the presented word (*Moses*) as the expected word (*Noah*) due to bottom-up (phonological) priming, top-down (semantic) priming, or both. Two experiments supported this miscomprehension account: Lexical- and proposition-level information contributed autonomously to miscomprehensions and Moses mistakes in Experiment 1, and prior presentation of nonanomalous information reduced subsequent anomaly detection in Experiment 2. Present results contradict accounts in which Moses mistakes involve semantic but not phonological processes, involve mechanisms different from everyday language comprehension, or involve special anomaly detection mechanisms for calculating the coherence between the Moses question and the anomalous word.

Keywords: Sentence processing; Comprehension; Moses illusion; Semantic processing; Phonological processing.

Why do people sometimes miscomprehend sentences that they believe they have comprehended correctly? This question is important for understanding human communication and education, and has motivated research on the "Moses illusion" and related comprehension mistakes in listening (e.g., Shafto & MacKay, 2000) and reading (e.g., Erickson & Mattson, 1981). To revisit the classic example, listeners who know that Noah rather than Moses

Correspondence should be addressed to Don MacKay, Psychology Dept., University of California, Los Angeles, Los Angeles CA 90095-1563, USA. E-mail: mackay@ucla.edu

The authors gratefully acknowledge support from the Samuel A. MacKay Memorial Research Fund and NIA grant R01AG 09755. Dr. Meredith A. Shafto is currently at the Centre for Speech, Language, and the Brain, Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK. The authors thank Dr. Michael Shafto for advice on statistics and experimental design, Dr. Deborah Burke for helpful comments on an earlier draft, and Irwin Stein for recruiting participants.

http://www.psypress.com/ecp DOI: 10.1080/09541440902941967

launched the Ark often fail to detect the anomaly in the question *How many animals of each kind did Moses take on the Ark?* and mistakenly respond *two.* This Moses illusion remains stable over time, as with visual illusions. For example, when tested after a delay, participants misremember as non-anomalous the anomalous *Moses* question that they saw, which indicates that they originally miscomprehended the anomalous question as nonanomalous (see Shafto & MacKay). Also as with visual illusions, people experiencing Moses illusions often feel that they "missed" or "misperceived" something obvious, e.g., the word *Moses.* However, methodological controls rule out purely perceptual explanations: Moses mistakes persist when participants accurately read the anomalous question aloud (Erickson & Mattson) or when they accurately shadow the substituted word (*Moses*) during auditory presentation (Shafto & MacKay).

A major theoretical issue associated with these "illusions" is whether they result from the same processes as comprehending error-free material, or whether they reflect failures in anomaly detection mechanisms that are separable from other language comprehension processes. The latter view was predominant in initial attempts to explain the Moses illusion, and the most prominent example of this approach is the Partial Match Hypothesis (PMH; Barton & Sanford, 1993; Kamas, Reder, & Ayers, 1996; Reder & Cleeremans, 1990; Reder & Kusbit, 1991; van Oostendorp & Kok, 1990). Under PMH, people compute a partial or incomplete semantic analysis of anomalous questions which is then used for comparison with stored memory schemas. This comparison process provides a measure of "conceptual cohesion" or "semantic coherence", which varies with the overlap in semantic features between the question and the stored information. If semantic overlap in the preliminary partial analysis is extensive, yielding a cohesion measure that exceeds some adjustable criterion, the system aborts the comparison process, and calls for an answer to the question, introducing a Moses mistake. However, if the cohesion measure falls below criterion, then more complete or detailed processing occurs that culminates in detection of the anomaly.

Under an alternative approach adopted in our earlier research (Shafto & MacKay, 2000), Moses mistakes are a consequence of general comprehension processes within an interactive theory of language and memory called Node Structure Theory (NST; e.g., MacKay, 1987). Under NST, Moses mistakes occur when people miscomprehend the substituted word (*Moses*) as the expected word (*Noah*), so that the likelihood of errors depends not on a checking procedure, but on the same processes as error-free comprehension, primarily top-down and bottom-up priming or partial activation (discussed shortly). Moreover, individual semantic features are not simply tallied under NST: Their contribution depends on interactions with other semantic and nonsemantic features and their locus within the hierarchically organised language system. NST shares a number of characteristics with other interactive models of language and memory, but, as the current research includes predictions specific to NST, we begin by outlining relevant details of this theory.

NST postulates two theoretical processes relevant to language comprehension and Moses mistakes: *priming* and *activation*. Following its original use in Lashley (1951), the term *priming* refers to below-threshold activity that prepares a representational unit or *node* for activation and under NST, priming spreads (with decrement) between connected nodes: As a theoretical concept, this *priming* or partial activation process is distinct from later uses of the term *priming* to designate empirical variables or tasks such as "repetition priming" or "semantic priming" (see MacKay, 1987, pp. 10–12). By contrast, *activation* is all-or-none and sequential, does not spread, and occurs when a selection mechanism activates the node with the most priming in a *domain* or activation category at some discrete point in time. This "most-primed-wins" activation principle is a primary cause of both everyday comprehension errors (see MacKay, 1987, pp. 62–89) and the Moses illusion: Under NST, Moses mistakes occur because semantic, orthographic, or phonological sources alone or in combination contribute more *priming* to the lexical node for *Moses* than for *Noah*, so that *Moses* becomes *activated* and comprehended under the most-primed-wins activation principle.

Under NST, shared phonology can alone suffice to induce Moses mistakes (see Shafto & MacKay, 2000). As a first step in addressing this prediction, Shafto and MacKay (2000) demonstrated an "Armstrong illusion" that resembled the Moses illusion except that the miscomprehensions resulted from extensive overlap in phonology between the expected and substituted names. More comprehension errors occurred with substitution of phonologically related names, e.g., *Louis Armstrong* for *Neil Armstrong* in the anomalous question *What was the famous line uttered by Louis Armstrong when he first set foot on the moon?*, than with substitution of otherwise similar but phonologically unrelated names (e.g., *Dizzy Gillespie* instead of *Louis Armstrong*).

Shafto and MacKay (2000) also demonstrated that more comprehension errors occurred with *combined* phonological and semantic overlap between an expected name (e.g., *Patrick Stewart*) and a substituted name (*Jimmy Stewart*), than with phonological overlap alone (*Rod Stewart*) or semantic overlap alone (*Paul Newman*). This additive effect of phonological and semantic overlap, known as the *mega-Moses effect*, is also consistent with the NST miscomprehension hypothesis.

To illustrate how semantic and phonological overlap combine to cause mega-Moses mistakes under NST, Figure 1 shows a subset of the nodes for comprehending the auditorily presented question, *Lead singer of The Doors, Van Morrison, is buried in what European city?* During comprehension of this

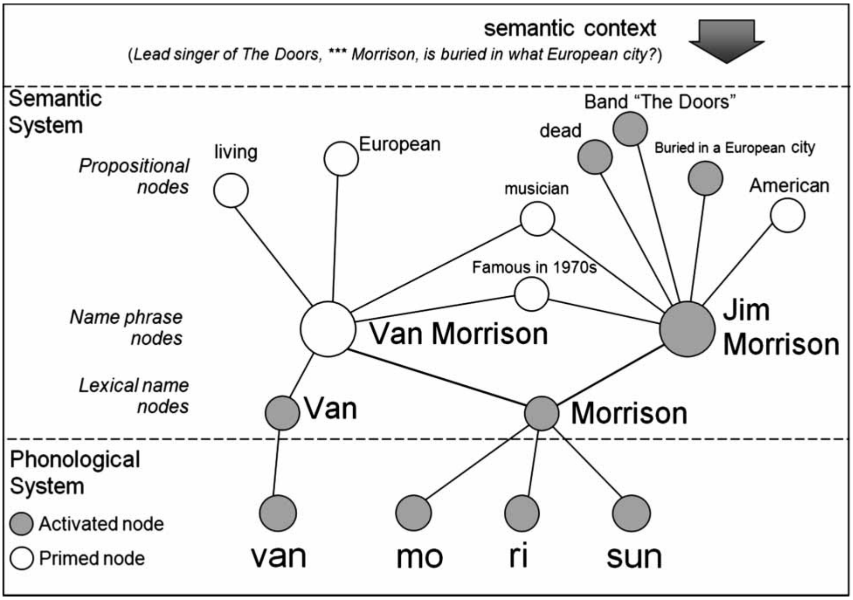

Figure 1. A subset of semantic system and phonological system nodes in NST and their two-way connections underlying miscomprehension of *Van Morrison* as *Jim Morrison* in the sentence, *Lead singer of The Doors, Van Morrison, is buried in what European city?* Grey nodes are activated, and white nodes are primed but not activated.

sentence under NST, phonological nodes for the proper name *Morrison* deliver equivalent levels of first-order bottom-up priming to the name phrase nodes for *Van Morrison* (currently living) and *Jim Morrison* (deceased). However, comprehension of the sentence context delivers additional, top-down priming to *Jim Morrison* for listeners who know about his death, his burial in Europe and his association with The Doors (see the semantic system links in Figure 1). The name *Jim Morrison* is therefore likely to receive more priming (summated from all sources) than *Van Morrison* and become activated in error under the most-primed-wins principle, causing miscomprehension of the anomalous name contained in this sentence (*Van Morrison*) as *Jim Morrison*.

GOALS OF THE PRESENT STUDY

The present study tested three predictions derived from the NST miscomprehension hypothesis. These predictions are relevant to issues raised by previous research (Shafto & MacKay, 2000) and fall under three general headings discussed next.

Effects of bottom-up phonological priming

Erickson and Mattson (1981) concluded that the Moses illusion is a strictly semantic phenomenon (despite the phonological overlap in stress pattern, bisyllabicity, initial vowels, onset-nasality, and onset-voicing of *Moses* and *Noah*) and the role of phonology has been largely ignored in research since then. Shafto and MacKay (2000) was an exception to this pattern and their Armstrong effect indicated that phonological overlap between the substituted and expected names can increase the frequency of Moses mistakes, consistent with the hypothesised role of bottom-up priming in NST. However, this result did not indicate that phonological overlap between the substituted and expected names can *alone* suffice to raise miscomprehension rates as per the NST miscomprehension hypothesis. To see why, note that the critical names *Louis Armstrong* and *Neil Armstrong* in the anomalous question, *What was the famous line uttered by Louis Armstrong when he first set foot on the moon?* are both twentieth-century figures, have appeared on television, are male, and share the same surname node within the semantic system. The resulting semantic-level overlap may therefore contribute to Armstrong mistakes, rendering the Armstrong effect a weak version of the mega-Moses effect illustrated in Figure 1, where the critical names *Van Morrison* and *Jim Morrison* likewise exhibit both phonological overlap *and* semantic overlap (albeit more extensive semantic overlap than just a shared lexical node).

To test for the *strictly phonological* Armstrong effect predicted under NST (see Shafto & MacKay, 2000), Experiment 1 in the present study eliminated all semantic overlap except for the shared surname node in Armstrong questions, and Experiment 2 went one step further by completely eliminating all semantic-level overlap (including the shared surname node). If residual semantic overlap is necessary for the occurrence of Armstrong miscomprehensions, then the reduced overlap in Experiment 1 should greatly reduce the Armstrong effect, and the nonexistent overlap in Experiment 2 should entirely eliminate the Armstrong effect. However, NST predicted a reliable Armstrong effect even when semantic overlap was nonexistent in Experiment 2.

Autonomous contributions from different systems

The present study focused on bottom-up phonological priming not just to test for the strictly phonological Armstrong effect predicted under NST, but to test alternative hypotheses as to the relation between semantic versus phonological processing systems and the relation between successes versus failures in anomaly detection. Although the PMH could in principle be extended to include the contribution of *phonological* feature matches (L. M.

Reder, personal communication, 2 September 2005), failure or success in anomaly detection would still depend on an overall coherence score that collapses the match–mismatch contributions from the semantic and phonological systems. In short, success and failure in anomaly detection are mirror image events under PMH, and so are the semantic and phonological contributions to this success or failure under PMH.

By contrast, the semantic and phonological systems can autonomously affect failure versus success in anomaly detection under NST. This is because, unlike priming, node activation is system-specific and category-specific under NST (see MacKay, 1987, pp. 39–61). As a result, the same stimulus can trigger activation of conflicting information within the phonological versus semantic systems. For example, the anomalous word *Louis* in the question *What was the famous line uttered by Louis Armstrong when he first set foot on the moon*? can cause activation of the phonological representation for Louis in the phonological system and the lexical node for Neil in the semantic system. To demonstrate this predicted phenomenon, Shafto and MacKay (2000) used a "partial shadowing" procedure: Unlike other anomaly detection studies where participants simply read or hear an anomalous question, participants in Shafto and MacKay heard the anomalous question while silently reading a written version of the question. The written version contained one or more blank slots and their task was to shadow or repeat aloud with minimal lag whatever auditory word occupied each slot. For example, participants shadowed the words Louis, Armstrong, and moon in the question, "What was the famous line uttered by Louis Armstrong when he first set foot on the moon?" The critical name (here, Louis Armstrong) was always among the shadowed words and we only scored responses to the question when participants correctly shadowed the critical words. When participants produced the word Louis during partial shadowing, this indicated activation of the phonological nodes for Louis under NST. Then, when the same participants answered the anomalous question as if it had contained Neil rather than Louis Armstrong, this indicated online activation and comprehension of the name Neil Armstrong instead of the presented name Louis Armstrong. The miscomprehension hypothesis was supported in a post-experimental recognition memory test, where participants experiencing the Armstrong illusion tended to misrecall that the word Neil had been presented instead of the actually presented Louis. Accurate shadowing followed by miscomprehension and misrecall therefore indicated that lower level activation can proceed autonomously relative to higher level activation in the Armstrong illusion and language comprehension in general (see MacKay, 1987, pp. 62–89).

The present study examined another way whereby nodes in different hierarchically organised categories can make autonomous contributions to anomaly detection under NST. The specific hypothesis under test was that lexical-level processes suffice for *failed* anomaly detection, whereas both lexical- and proposition-level processes are necessary for *successful* anomaly detection under NST. The role of propositional representations in anomaly detection has been under-examined, with research on anomaly detection typically focusing on the influence of lexical and sublexical processes (see, e.g., MacKay, 1972, Shafto & MacKay, 2000) rather than proposition-level processes. However, the process of anomaly detection differs for words in isolation versus in sentence contexts under NST, so that lexical factors do not completely determine Moses mistakes: When the expected name becomes activated in error instead of the substituted name, subsequent proposition-level processes are necessary to trigger detection that the sentence is anomalous (see Shafto & MacKay).

To illustrate why proposition-level processes are essential for successful anomaly detection, consider one of the anomalous sentences in Experiment 2: *In gorilla culture, the dominant mail must defend his status* (see Figure 2).

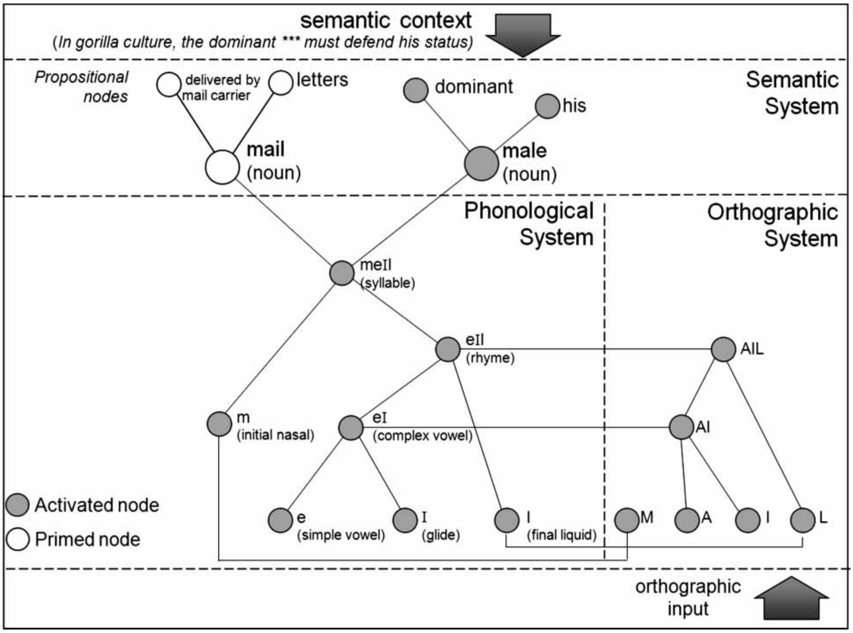

Figure 2. A subset of semantic, phonological, and orthographic system nodes in NST and their two-way connections underlying miscomprehension of *mail* as *male* in the anomalous sentence, *In gorilla culture, the dominant mail must defend his status.* Grey nodes are activated, and white nodes are primed but not activated. The theoretical and empirical bases for the connections shown between orthographic and phonological nodes are discussed in MacKay and James (2002). Connections linking orthographic system nodes representing M and AIL to the semantic system node representing *mail* (noun) have been omitted for simplicity.

Assume that the lexical node for *mail* becomes activated in this sentence context, yielding accurate comprehension of the semantically anomalous word *mail*. In order to integrate this lexical unit into a proposition that incorporates the remainder of this sentence, participants must imagine the existence of a new type of *mail* that is compatible with the everyday concept of postal *mail* and at the same time somehow designates a male animal (indicated via the pronoun *his* in the sentence context) that has defensible status in gorilla culture. However, it is difficult to imagine what this type of *mail* might refer to, and this proposition-level comprehension difficulty provides the signal that this sentence is anomalous. Thus, success and failure at anomaly detection are not entirely codependent or mirror image events under NST because, although lexical-level processes suffice to explain miscomprehension of the anomalous word and failed anomaly detection, an additional process is necessary to explain successful anomaly detection: proposition-level comprehension difficulty, which signals that the overall sentence is anomalous.

The substituted and expected names in Experiment 1 had low to zero proposition-level overlap, and this allowed us to test the possibility that lexical-level miscomprehension can occur *without* the proposition-level difficulties that trigger anomaly detection. We illustrate this possibility via an anomalous sentence with low proposition-level overlap from Experiment 1: *Although he has holdings in entertainment and the media, Nicholas Turner also owns what sports team?* Because *Nicholas Turner* is not a previously encountered name and has no internal representation except for maleness in semantic memory, *Nicholas Turner* lacks affiliated propositional information that is incongruous with the rest of the sentence. However, participants familiar with *Ted Turner* and his various holdings can integrate the substituted name with the sentence context by, for example, assuming that *Nicholas* is Ted's middle name, so that comprehension can proceed normally without the proposition-level difficulties that enable detection that the sentence is anomalous under NST. We label this hypothetical outcome the "unknown Armstrong effect". Participants experiencing this predicted effect will comprehend the unknown substituted name as the expected name under NST, just as in the standard Armstrong effect. However, they will not detect that the sentence containing the unknown name is anomalous because *inferred* proposition-level information prevents the proposition-level difficulty that triggers anomaly detection under NST.

Effects of prior exposure on anomaly detection

Experiment 1 tested the NST hypothesis that anomaly detection will decline with a lack of relevant contradictory propositional information. For the more typical Moses illusion materials, both the expected and presented names are affiliated with conflicting propositional information. Experiment 2 manipulated the relative *availability* of this conflicting information using a repeated-exposure paradigm. Under NST, exposure to information strengthens the connections representing that information, and reexposure facilitates subsequent activation via those same strengthened connections, increasing the likelihood of reinstating the original comprehension when the same material is presented again. This principle predicts an increase in Armstrong mistakes for the same anomalous sentence presented with versus without prior exposure to the *nonanomalous* version of that sentence. The reason is that prior exposure to the nonanomalous sentence strengthens the connections between the expected name and the sentence context. These strengthened connections in turn increase the likelihood of activating the expected name in error when the anomalous sentence is subsequently presented. Experiment 2 examined whether prior exposure to nonanomalous sentences yields this predicted increase in Armstrong mistakes for anomalous sentences.

EXPERIMENT 1: UNKNOWN ARMSTRONG EFFECTS

Experiment 1 had two specific aims. One was to replicate the original Armstrong effect (Shafto & MacKay, 2000) using a larger number of general knowledge questions in three conditions: the *nonanomalous* condition containing the expected name, the *Armstrong* condition containing a phonologically related name, and the *control* condition, containing a phonologically unrelated name. Questions in these conditions differed only in their *critical names*, e.g., *During which decade did Henry Ford (Gerald Ford, Herbert Hoover) introduce his Model T to the world?* Note that Armstrong names (e.g., *Gerald Ford*) and control names (e.g., *Herbert Hoover*) were similar in meaning (both Gerald Ford and Herbert Hoover were US presidents) but differed in phonology and orthography.

The second specific aim of Experiment 1 was to test for the existence of unknown Armstrong effects predicted under NST. For this test, two types of unknown names were substituted for the expected names in anomalous questions: *unknown same-gender* names and *unknown different-gender* names. The unknown same-gender names were nonfamous, unknown to the participants, and of identical gender to the nonanomalous name expected on the basis of sentence context. For example, *Michael Armstrong* was the unknown same-gender name substituted for *Neil Armstrong* in questions such as *What was the famous line uttered by Michael Armstrong when he first set foot on the moon?* The unknown different-gender names were similar but opposite in gender from the expected or nonanomalous name. For example, *Mary Armstrong* was the unknown different-gender name substituted for the expected *Neil Armstrong*. Note that an unknown same-gender name reduces the residual propositional information shared with the nonanomalous name, and an unknown different-gender name reduces the shared semantic information even further by eliminating the overlap in gender. Results of Experiment 1 will therefore determine whether the presence of residual semantics (e.g., shared gender information between the anomalous and expected words) is sufficient to explain the standard Armstrong effect: If residual shared semantics determines the standard Armstrong effect, this effect should diminish or disappear in the unknown name conditions. However, the NST miscomprehension hypothesis predicted elevated miscomprehension rates compared to the control condition in both unknown name conditions because of the shared phonology between the nonanomalous and unknown names.

The presence or absence of conflicting gender information in unknown name conditions also allowed a test for proposition-level effects predicted under the NST miscomprehension hypothesis: We expected more anomaly detection in the unknown different-gender condition than in the unknown same-gender condition because the conflicting gender information in the different-gender condition can cause a proposition-level breakdown in comprehension. However, both types of unknown names reduce proposition-level conflict relative to known names because little or no propositional knowledge associated with the unknown names conflicts with the sentence context. Thus, even though the unknown names reduce shared semantic overlap, they also reduce the semantic conflicts for enabling detection that the overall sentence is anomalous, and we expected a relatively high rate of Armstrong mistakes in unknown name conditions.

As in Shafto and MacKay (2000), Experiments 1 and 2 included a "comprehension-memory" test that tapped comprehension at a delay. Using a procedure developed in MacKay (1973), questions from the initial anomaly detection task were reworded so as to either preserve or alter their original meaning. Then, in the offline recognition-memory test, participants saw the reworded questions and indicated whether each was the "same" or "different" in meaning "for even a single word" from a question previously presented during the anomaly detection phase. In fact, the reworded question always had the same meaning as the *nonanomalous* version of the experimental question, and the correct response was "different" when *anomalous* versions of the question were presented in the anomaly detection phase. For example, the anomalous question, *What was the famous line uttered by Michael Armstrong when he first set foot on the moon?* was reworded as, *When he first set foot on the moon, Neil Armstrong uttered what famous line?* for the offline recognition-memory test. Under NST, participants should produce more "same" responses to comprehension-memory questions after they fail to detect the anomaly and answer the anomalous question than after they detect the anomaly and correctly respond "can't say" in the anomaly detection phase. The reason is that answering an anomalous question indicates miscomprehension of the anomalous name as the nonanomalous name under NST, so that participants will be more likely to "remember" the nonanomalous name in the comprehension-memory task. Thus, more "same" responses should occur for Armstrong than for completely unrelated or control questions due to the increased miscomprehension of Armstrong names as nonanomalous names in the anomaly detection phase. Additionally, more "same" responses should occur for unknown same-gender than unknown different-gender and control questions due to the increased miscomprehension of unknown same-gender names as nonanomalous names in the anomaly detection phase.

Method

Participants

Participants were 30 UCLA undergraduates aged 18–22 ($M = 19.3$) who participated for partial course credit in an introductory psychology course.

Materials

Materials were 80 general knowledge questions: 40 experimental and 40 filler questions. The experimental questions were 8–23 words long ($M = 15.60$, $SD = 3.02$) and always contained a *critical name*, consisting of a first and last proper name. Each experimental question came in five versions: the nonanomalous version containing the expected name, and four anomalous versions. The anomalous versions contained critical names that were either related or unrelated in phonology to the nonanomalous name. Two anomalous versions were designed to replicate the Armstrong effect: the *Armstrong* condition containing a critical name phonologically related to the nonanomalous name, and the phonologically unrelated *control* condition. The remaining two anomalous versions were the *unknown same-gender* and *unknown different-gender* conditions described earlier (see Table 1 for examples).

Each question had three multiple choice answers: the answer to the nonanomalous version of the question, an incorrect but nevertheless reasonable answer to the nonanomalous version for the question, and "can't say", which was the correct answer to anomalous versions (see Table 2 for examples). To illustrate, the answer to the Armstrong version, *During which decade did Gerald Ford introduce his Model T to the world?* was "can't say", the correct answer for the nonanomalous version, *During which decade did Henry Ford introduce his Model T to the world?* was "1900s", and the incorrect but reasonable alternative for this nonanomalous version was "1940s".

The filler questions were either nonanomalous ($N = 20$) or anomalous ($N = 20$) and served to dissuade a focus on proper names and to ensure

TABLE 1
Example critical names, auditory sentences, corresponding text for partial shadowing, and results for the anomaly detection phase in Experiment 1

Condition	*Critical name*	*Example auditory stimulus*	*Text for silent reading (with blanks for partial shadowing)*	*Mean number of "can't say" responses per participant*	*Mean number of valid substantive responses per participant*
Experimental questions					
Nonanomalous	Henry Ford	During which decade did Henry Ford introduce his Model T to the world?	During which decade did _______ _______ introduce his Model T _______ the world?	1.40 (1.52)	5.03 (1.69)
Armstrong	Gerald Ford	During which decade did Gerald Ford introduce his Model T to the world?	During which decade did _______ _______ introduce his Model T _______ the world?	5.50 (1.98)	1.20 (1.21)
Control	Herbert Hoover	During which decade did Herbert Hoover introduce his Model T to the world?	During which decade did _______ _______ introduce his Model T _______ the world?	5.73 (1.84)	0.40 (0.56)
Unknown same-gender	William Ford	During which decade did William Ford introduce his Model T to the world?	During which decade did _______ _______ introduce his Model T _______ the world?	4.27 (2.35)	2.40 (2.09)
Unknown different-gender	Laura Ford	During which decade did Laura Ford introduce his Model T to the world?	During which decade did _______ _______ introduce his Model T _______ the world?	4.93 (2.18)	1.13 (1.53)
Filler questions					
Nonanomalous	—	In the biblical story, who was swallowed by the whale?	In the biblical _______, who was swallowed _______ the whale?	1.03 (1.43)	18.00 (2.72)
Anomalous	—	Which planet in our television system is closest to the Sun?	Which _______ in our television _______ is closest to the Sun?	17.60 (1.83)	0.00 (0.00)

TABLE 2
Example questions and answer choices for a filler and Armstrong question in the three phases of Experiment 1

Phase	*Questions*	*Answer choices*
Experimental questions		
Anomaly detection phase	During which decade did Gerald Ford introduce his Model T to the world?	a. *CAN'T SAY* b. 1940s c. 1900s
Comprehension-memory phase	Henry Ford introduced his Model T to the world during which decade?	"same" "*different*"
Knowledge-verification phase	Who introduced his Model T to the world in the early 1900s?	a. William Ford b. Herbert Hoover c. *Henry Ford* d. Lee Iacocca e. Gerald Ford f. Laura Ford
Filler questions		
Anomaly detection phase	How many letters are there in the alphabet?	a. CAN'T SAY b. *26* c. 31
Comprehension-memory phase	There are how many numbers in the alphabet?	"same" "*different*"

Correct answer choices are italicised.

correct use of the "can't say" and substantive response alternatives. To dissuade a proper name focus, anomalous fillers never contained anomalous proper names, e.g., *How many numbers* (*letters*) *are there in the alphabet?* (see also Table 1). To ensure correct use of the "can't say" response, anomalies were intentionally easy to detect in anomalous fillers, e.g., *In what mythology was Venus considered the god of Computers?* To ensure correct use of substantive response alternatives, nonanomalous fillers, e.g., *What American city is known as "The Big Apple"?* had answers that were intentionally easy to recognise (see also Table 2).

In constructing the final versions of our experimental stimuli, we ensured that the question context uniquely typified the nonanomalous names, that participants were likely to answer nonanomalous and filler questions correctly, and that familiarity with the famous critical names was high across the four conditions in Experiment 1. To do this, we "filtered" a large number of draft stimuli through six "filter studies" with between 10 and 32 participants per study. In Filter Study 1, participants rated their familiarity with 412 famous names and indicated their "reasons for fame". We filtered out names with few correctly identified reasons for fame, and used the remaining names to construct 40 name sets, each with a nonanomalous, Armstrong, and control name. These name sets were used to construct name pairs for comparing the four conditions, and new participants in Filter Study 2 then rated these name pairs for semantic relatedness. Results confirmed that Armstrong names were more semantically related to control names than either Armstrong or control names were to the nonanomalous names. We next incorporated the nonanomalous names into 40 experimental questions that we presented to participants in Filter Study 3. Results of Filter Study 3 confirmed that participants were able to answer the nonanomalous form of the questions with high probability. To verify that participants had sufficient knowledge to differentiate the nonanomalous name from the anomalous names in the context of each question, participants in Filter Study 4 performed a five-choice recognition memory test involving the nonanomalous, Armstrong, and control names. For example, response alternatives for, *Who said "...One small step for man ..." when he first set foot on the moon?* were: Louis Armstrong, Duke Ellington, Alan Shepard, Mary Armstrong, Jacob Armstrong, and Neil Armstrong. In order to ensure that unknown names were unfamiliar, new participants in Filter Study 5 rated these names for "degree of fame". Finally, participants in Filter Study 6 answered the filler questions, with results indicating that nonanomalous filler questions were readily answered, and that the anomalies in anomalous filler questions were readily recognised.

Procedure

Participants fitted with headphones and a small lapel microphone sat facing a computer monitor that presented instructions subsequently summarised by the experimenter. As in Shafto and MacKay (2000, Exp. 2), the computer presented general knowledge questions in three phases: anomaly detection, comprehension-memory, and knowledge-verification.

Anomaly detection phase. Participants saw 80 general knowledge questions in a multiple-choice answer format. Participants were warned that some of the questions would contain *anomalies*, i.e., information that conflicted with the remainder of the question. For questions containing anomalies, participants were instructed to choose the "can't say" response. They received example questions to illustrate correct "can't say" responses, followed by 10 practice trials. The experimenter next reiterated the instructions, answered any procedural questions, and began the 80 anomaly detection trials.

Each trial involved *partial shadowing* followed by *anomaly detection*. During partial shadowing participants heard a question over the headphones while silently reading a written version that contained one to three blank slots. Participants shadowed or repeated aloud with minimal lag the auditory word occupying each slot. For example, for the question, *During which decade did Gerald Ford introduce his Model T to the world?*, participants shadowed the words *Gerald*, *Ford*, and *to* (see Table 1). The shadowed words in experimental questions always included the critical names (first name and surname) to ensure that misperception or failure to attend to the critical names could not explain our results. Participants' shadowing responses were recorded on an answer sheet by the experimenter, and recorded on audio tape for subsequent checking. Following partial shadowing, participants pressed the space bar, which triggered the anomaly detection task. Participants selected the correct answer from the three multiple-choice alternatives including "can't say" and two substantive answers.

Comprehension-memory phase. The comprehension-memory phase was a surprise test of recognition memory for the *meaning* of anomaly detection questions. Instructions informed participants that they would see reworded versions of anomaly detection questions that either preserved or distorted the meaning of original questions. They would respond "same" via keypress if the reworded question was the same *in meaning* to its prior auditory version, and responded "different" otherwise.

Following the instructions, participants saw the 80 reworded experimental and filler questions presented visually in the same order as during anomaly detection. Word order always differed for these recognition targets versus the corresponding anomaly detection questions. For example, word order differs in the anomaly detection question *During which decade did Gerald Ford introduce his Model T to the world?* versus in the recognition question *Henry Ford introduced his Model T to the world during which decade?* As in this example, all reworded experimental questions were synonymous with the nonanomalous version of the corresponding anomaly detection question, so that the correct response was "different" for all anomalous questions and "same" only for nonanomalous questions. To mask the relative infrequency of correct "same" responses to experimental questions, all anomalous fillers remained anomalous and thus "same" when reworded, and all nonanomalous fillers became anomalous and thus "different" when reworded.

Each comprehension-memory question appeared centred on the screen until the participant responded "same" or "different". The question "CONFIDENCE?" and a 0–4 scale then appeared, and participants estimated their confidence in their decision via keypress, and the trial advanced automatically.

Knowledge-verification phase. The knowledge-verification phase ensured that participants had the knowledge necessary for responding correctly during anomaly detection. Instructions informed participants that they would answer a series of multiple-choice questions, and participants then saw 40 general knowledge questions about the nonanomalous names in experimental questions (see the example in Table 2). Each question had six proper names as response alternatives: the Armstrong, control, unknown names, a semantically plausible but incorrect famous name, and the nonanomalous name, which was always the correct answer. After participants responded via keypress, the 0–4 scale and "CONFIDENCE?" question reappeared, and participants rated confidence in their response.

Results and discussion

Before all data analyses, we removed invalid responses. These included trials with errors in shadowing the critical name (1.25% of all trials), invalid keypresses (1.58% of trials), and inaccurately answered knowledge-verification questions (10.58% of all trials). Thus, we only analysed trials where participants had accurate knowledge about the critical names and perceived them correctly in the anomaly detection phase. All statistics were two-tailed except for predicted effects noted otherwise. Table 3 shows for all five conditions the mean proportion of "can't say" responses to general knowledge questions in the anomaly detection phase, together with the mean proportion of "same" responses in the comprehension-memory phase. Nonparametric Friedman ANOVA tests were used to test for main effects across conditions, followed by Wilcoxon tests for planned comparisons between conditions. Results of analyses by subject and by item are reported for each comparison.

Anomaly detection results

Figure 3 shows our measure of anomaly detection, the mean proportion of "can't say" responses in the anomaly detection phase for the standard Armstrong conditions (control and Armstrong names) and the unknown conditions. There was an overall main effect of condition, Friedman ANOVA test, $p < .001$ by subject, $p < .001$ by item, with a lower "can't say" proportion for Armstrong names than control names, Wilcoxon $z = 2.03$, $p < .05$ by subject, $z = 2.60$, $p < .01$ by item, an effect that indicates replication of the standard Armstrong effect (Shafto & MacKay, 2000).[1]

[1] This effect is relevant to the original Moses question in Erickson and Mattson (1981) because of the low-level phonological similarity between their critical words, *Moses* and *Noah* (which overlap in stress pattern, bisyllabicity, initial vowels, onset-nasality, and onset-voicing). Because the low-level phonological overlap between the phonologically similar and nonanomalous words in the present study did not affect anomaly detection, this suggests that Erickson and Mattson's results involve a simple Moses effect rather than a mega-Moses effect. However, caution is warranted on this issue. If low-level phonological overlap has weak but reliable effects for "online" anomaly detection responses during sentence presentation (as in Erickson and Mattson), then Erickson and Mattson's discovery is more appropriately characterised as a mega-Moses effect (see Shafto & MacKay, 2000).

TABLE 3
Experiment 1 results: Mean proportion "can't say" responses in the anomaly detection phase, and mean proportion "same" responses in the comprehension-memory phase (*SD*s in parentheses)

	Standard Armstrong conditions			*Unknown conditions*	
	Nonanomalous	*Armstrong*	*Control*	*Same-gender*	*Different-gender*
Anomaly detection "can't say" responses	.18 (0.20)	.78 (0.22)	.87 (0.20)	.60 (0.31)	.73 (0.27)
Comprehension-memory "same" responses	.94 (0.11)	.19 (0.18)	.10 (.14)	.19 (0.16)	.10 (0.13)

Other planned comparisons revealed lower "can't say" proportions for unknown same-gender names than control names, Wilcoxon $z = 3.59$, $p < .001$ by subject, $z = 4.80$, $p < .001$ by item, and for unknown different-gender names than control names, Wilcoxon $z = 2.38$, $p < .05$ by subject, $z = 3.47$, $p < .01$ by item, indicating that both unknown name conditions yielded unknown Armstrong effects (see Figure 3). In addition, the "can't say" proportion was lower for unknown same-gender names than for Armstrong names, Wilcoxon $z = 3.44$, $p < .01$ by subject, $z = 2.88$, $p < .01$ by item, and for unknown same-gender than for unknown different-gender names, Wilcoxon $z = 2.40$, $p < .05$ by subject, $z = 2.84$, $p < .01$ by item, indicating a larger unknown Armstrong effect for unknown same-gender names than for either the Armstrong names or the unknown different-gender names. The "can't say" proportion did not differ for unknown different-gender names versus Armstrong names, $p > .10$ by subject and by item, suggesting that, despite the gender conflict, unknown different-gender names yielded an unknown Armstrong effect that was as large as the standard Armstrong effect.

Anomaly detection results for Experiment 1 supported NST predictions. NST predicted a standard Armstrong effect, i.e., greater anomaly detection for control than Armstrong names. Also consistent with present results (see Figure 3), NST predicted less anomaly detection for unknown same-gender names than for unknown different-gender and control names, and less anomaly detection for unknown different-gender names than for control names.

Comprehension-memory results

For the comprehension-memory phase, we first compared the mean proportion of "same" responses for participants who had made correct ("can't say") responses versus incorrect (substantive) responses to anomalous

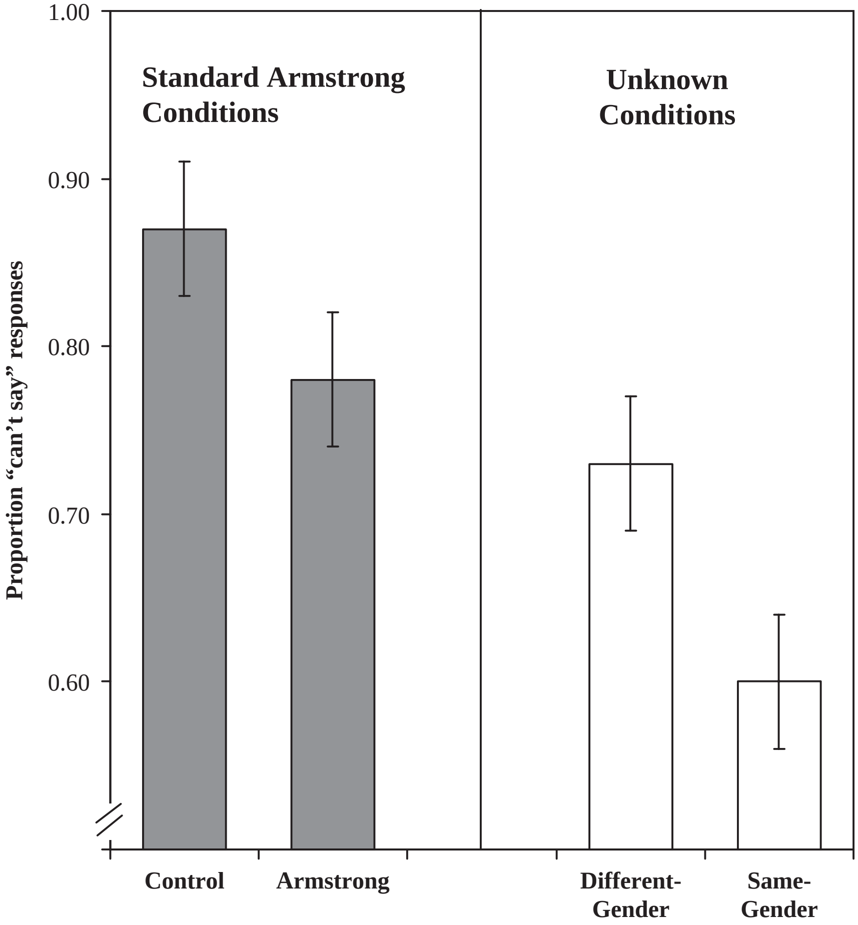

Figure 3. Mean proportion of "can't say" responses for the standard Armstrong conditions (left panel) and unknown conditions (right panel) in the anomaly detection phase of Experiment 1. The error bars indicate ±1 SE.

questions in the anomaly detection phase. The results indicated reliably more "same" responses following incorrect than correct responses, Wilcoxon $z = 4.40$, $p < .001$ by subject, $z = 4.66$, $p < .001$ by item. Because the critical name in comprehension-memory stimuli was always the nonanomalous name, this result indicates that when participants detected the anomaly during the anomaly detection phase, they miscomprehended (and later misrecalled) the nonanomalous name less often than if they failed to detect the anomaly.

We next examined the overall mean proportion of "same" responses, which is shown by condition in Table 3. A Friedman ANOVA comparing the anomalous (control, Armstrong, and unknown) conditions yielded a main effect, $p < .05$ by subject and by item, which reflected several differences: a larger "same" proportion for Armstrong names than control names, Wilcoxon $z = 2.51$, $p < .05$ by subject, $z = 2.20$, $p < .05$ by item, mirroring the standard Armstrong effect in anomaly detection; a larger "same" proportion for unknown same-gender names than control names, Wilcoxon $z = 2.37$, $p < .05$ by subject, $z = 2.51$, $p < .05$ by item, and for unknown same-gender names than unknown different-gender names, Wilcoxon $z = 2.65$, $p < .01$ by subject, $z = 2.81$, $p < .01$ by item, but no significant differences between unknown same-gender names and Armstrong names or between unknown different-gender names and control names, largest $z < 1$.

Taken together, the results of Experiment 1 support the prediction that different hierarchical levels contribute to Armstrong effects during sentence comprehension: Activation of an incorrect or nonpresented lexical representation can lead to miscomprehension, but anomaly detection will nonetheless remain low if little or no conflicting proposition-level information is activated (as in the unknown name conditions). The unknown name conditions also provided further support for bottom-up phonological priming as a contributor to these errors.

In sum, Experiment 1 results support two predictions derived from the NST approach to the Moses illusion: first, that bottom-up phonological priming can lead to the erroneous activation and comprehension of expected but unpresented names within the semantic system, and second, that such miscomprehensions will not be detected when the propositional information that becomes activated is consistent with the expected but unpresented name. These findings are problematic for feature matching accounts like the PMH, which have not addressed the role of bottom-up priming. More fundamentally, the PMH and related models would predict *high* anomaly detection rates in unknown name conditions, where feature overlap with the sentence is low. However, we found *low* anomaly detection rates in unknown name conditions, with participants detecting same-gender unknown names less often than standard Armstrong illusion names.

Nonetheless, limitations of the present data must be stressed. First, the unknown names in Experiment 1 reduced information inconsistent with the nonanomalous sentence context but did not directly manipulate the availability of propositional information *consistent* with this context, the primary driver of Moses mistakes under NST. To overcome this limitation, Experiment 2 employed a reexposure paradigm that directly increased the availability of this "consistent" propositional information. Second, Experiment 1 did not rule out the possibility that Armstrong effects reflect shared semantic information at least in part because the unknown and Armstrong names shared a lexical node in the semantic system (e.g., the *Morrison* node in *Figure 1*). To rule out this possibility and test for the strictly phonological

Armstrong effect predicted under NST, Experiment 2 used critical words that were not proper names.

EXPERIMENT 2: PHONOLOGICAL ARMSTRONG EFFECTS

Experiment 2 had two specific aims: to determine whether the shared lexical node plays an essential role in Armstrong mistakes, and to examine how prior auditory exposure to nonanomalous sentences affects the subsequent detection of anomalies in anomalous sentences.

In the prototypical Armstrong question *What was the famous line uttered by Louis Armstrong when he first set foot on the moon?*, *Louis Armstrong* and *Neil Armstrong* share a lexical node, namely the surname node (*Armstrong*). As a consequence, the standard Armstrong effects in Shafto and MacKay (2000) and Experiment 1 may reflect this shared semantic component instead of or in addition to the phonological components shared at lower levels. However, this possibility cannot be tested using proper names in standard Moses or Armstrong tasks because a set of surname pairs with identical phonology but no shared lexical surname node does not exist. We therefore created critical words from classes other than proper names in Experiment 2, and we eliminated semantic overlap between phonologically identical nonanomalous and anomalous names by using homophone pairs. For example, the critical words underlined in the sentence, *In gorilla culture, the dominant male* (*mail*, *mill*, *shed*) *must defend his status* are common rather than proper nouns and the (underlined) word substitutions in the anomalous conditions (*mail*, *mill*, *shed*) share no semantic units (e.g., propositional, phrase, or lexical nodes) with the nonanomalous critical word (*male*). As a consequence, differing results for the three anomalous word conditions can only reflect phonological overlap. Note also that the word substitutions vary systematically in their degree of phonological overlap with the nonanomalous word: phonologically unrelated for the *control* word, *shed*; *phonologically similar* for *mill*; and *phonologically identical* for *mail*.

NST predicted a nonmonotonic relation between anomaly detection and word condition in Experiment 2, with less anomaly detection for the phonologically identical than control words, but little or no difference between the phonologically similar versus control words. Figure 2 illustrates the basis for these predictions for the sentence, *In gorilla culture, the dominant mail must defend his status.* As in the standard Armstrong effect, the lexical node for the critical word (*male*) receives both top-down semantic priming from the sentence context and bottom-up phonological priming from the *mail/male* syllable node, which connects to both *male* and *mail* in the semantic system. As in the standard Armstrong effect, these sources of priming converge to activate the *male* node in error and cause miscomprehension of *mail* as *male* under NST.

However, unlike the phonologically identical word (*mail*), the phonologically similar word (*mill*) has no direct bottom-up syllabic connection to the lexical node for the nonanomalous word (*male*), so that bottom-up priming cannot induce miscomprehensions in the phonologically similar condition in the same manner as in the phonologically identical condition. Similarly, the syllable *shed* in the control condition has no direct bottom-up connection to the lexical node for *male*, so that even though *mill* enjoys greater phonological and orthographic overlap with *male* than *shed*, neither *mill* nor *shed* shares a syllable node with *male*, making it less likely that bottom-up priming will reliably induce miscomprehensions in either condition. As a consequence, NST predicts no more anomaly detection in the phonologically similar condition (*mill-male*) than in the control condition (*shed-male*).

The second goal of Experiment 2 was to examine how prior auditory exposure to the nonanomalous sentences affects subsequent anomaly detection. To manipulate prior auditory exposure, participants in Experiment 2 shadowed auditorily presented nonanomalous sentences, and then in the subsequent anomaly detection phase, they *saw* sentences that were either the same or different in meaning from the ones they shadowed previously.

NST predicts decreased anomaly detection for all anomalous versions of exposed sentences, because prior auditory processing of the nonanomalous sentence (e.g., *In gorilla culture, the dominant male must defend his status*) will strengthen top-down semantic connections to the word expected on the basis of sentence context (*male*). As a result, the sentence context will transmit more priming to the nonanomalous or expected word for exposed than unexposed sentences in all three anomalous conditions (phonologically identical, phonologically similar, and control).

Additionally, NST predicts a differential effect of word type for sentences which were previously exposed because of the direct syllable-level connection between phonologically identical words (*mail*) and nonanomalous words (*male*) discussed above. As can be seen in Figure 2, prior processing of the nonanomalous noun phrase (*the dominant male*) will strengthen the bottom-up phonological connections from the *male-mail* syllable to *male* in the phonologically identical condition. During subsequent presentation of the phonologically identical but anomalous sentence, *In gorilla culture, the dominant mail must defend his status*, these strengthened bottom-up connections to *male* will deliver enhanced bottom-up priming that increases the probability of activating the *male* node in error, causing miscomprehension of *mail* as *male* and reducing the probability of detecting the anomalous *mail*-for-*male* substitution. However, no direct connection links the phonological syllable for either *mill* or *shed* to the lexical node for *male*, ruling out the possibility of exposure-enhanced bottom-up priming of *male* via *mill* or *shed* (see Figure 2).

In sum, the sentence comprehension task in Experiment 2 tested assumptions about bottom-up phonological priming, the structure of phonological representations, and the strengthening of phonological–propositional connections via exposure. Our paradigm resembles paradigms in research on models of reading, especially paradigms examining whether phonological recoding operates during first pass processing of isolated words (e.g., Van Orden, 1987; see Brysbaert, Grondelaersb, & Ratinckxa, 2000, for a review) or during working memory processing (e.g., Waters, Caplan, & Leonard, 1992). It is not the goal of Experiment 2 to test between alternative reading models, but such similarities help to further our goal of relating the Moses and Armstrong illusions to general models of sentence reading and comprehension.

General procedures and summary predictions for Experiment 2

As in Experiment 1, participants answered two questions in the test phase of Experiment 2: an anomaly detection question and a comprehension-memory question. The anomaly detection question directly tapped anomaly detection: *Was that a valid sentence?* This direct measure improved on general knowledge questions, the indirect measure of anomaly detection used in Experiment 1 and most other studies of Moses mistakes (see Reder & Cleeremans, 1990, and Reder & Kusbit, 1991, for a discussion of problems associated with the question answering paradigm).

The second question in the test phase of Experiment 2 measured comprehension memory: *Did you hear that exact sentence before?* Procedures for this comprehension-memory question resembled Experiment 1 except that Experiment 2 participants received the comprehension-memory question immediately after each sentence in the test phase, and responded "yes" if they considered the test sentence completely identical *in meaning* to a sentence from the prior (auditory exposure) phase. Because only nonanomalous test sentences were completely identical in meaning to the prior (exposed) auditory sentence, "no" responses to anomalous test sentences indicated detection that the sentence is anomalous under NST, whereas "yes" responses indicated miscomprehension of the anomalous word as the nonanomalous word. NST therefore predicted the same nonmonotonic pattern for comprehension-memory questions as for preexposed anomaly detection questions: equal "yes" responses for phonologically similar and control versions, but more "yes" responses to phonologically identical than phonologically similar versions, reflecting the especially high probability of miscomprehending phonologically identical words (*mail*) as nonanomalous words (*male*) due to the exposure-linked increase in bottom-up priming of the nonanomalous words.

Method

Participants

Participants were 32 UCLA undergraduates who participated for partial course credit.

Materials

Materials were 40 experimental and 20 filler sentences of similar length ($M = 9.85$ words, $SD = 1.09$). Experimental sentences came in a nonanomalous and three anomalous versions created via substitution of a single critical word, e.g., *Jessica had to pay her* *bail* (*bale*, *boil*, *tank*) *before leaving the police station* (critical words underlined). The anomalous versions were labelled phonologically identical, e.g., *bale*; phonologically similar, e.g., *boil*; and control (phonologically unrelated), e.g., *tank* (see Table 4 for additional examples).

The four types of critical words were selected from Coltheart (1981) to have different meaning, but identical length in letters and syllables, and equivalent mean word frequency ($M = 36.34$, $SD = 46.44$, $R = 1$–275), with no significant differences between mean word frequency in any two pairs of conditions (all $ps > .10$). See the Appendix for a full list of our critical target words. Phonologically similar and phonologically identical words were also equally similar to the nonanomalous words in orthography, but not phonology. Orthographic similarity (calculated as in Van Orden, 1987) did not differ reliably for phonologically identical versus phonologically similar words, $p > .10$, ruling out orthographic explanations of differences between those conditions. However, phonological overlap (defined as the number of identical phonemes in the same position, according to the CMU Pronouncing Dictionary, 1998) did differ reliably for phonologically identical versus phonologically similar words, $t(78) = 14.00$, $p < .001$, so that differences in phonology rather than orthography characterise the phonologically identical versus phonologically similar conditions. Moreover, phonological overlap between nonanomalous and control words was low (4.8%), indicating that the nonanomalous and control words were indeed unrelated in phonology. By contrast, phonological overlap with the nonanomalous words was high (100%) for the phonologically identical words, and moderate (64%) for the phonologically similar words.

To ensure that participants could correctly spell the critical words in the nonanomalous condition, 20 pilot participants listened over headphones to 48 nonanomalous sentences containing nonanomalous words that were contextually unambiguous and wrote each sentence on a response sheet

TABLE 4
Example nonanomalous sentences and critical words in the test phase of Experiment 2

Critical words in the test phase by condition				
Nonanomalous	*Phonologically identical*	*Phonologically similar*	*Control*	*Example nonanomalous sentences (critical word italicised)*
DEER	DEAR	DOUR	LAME	As the truck passed the field, *deer* went running.
MALE	MAIL	MILL	SHED	In gorilla culture, the dominant *male* must defend his status.
PAIR	PARE	PAIN	WAIT	Everyone said what a lovely *pair* the bride and groom made.
POLE	POLL	PILE	HARM	Jimmy lost a ski *pole* and a boot when he fell.
REEL	REAL	RAIL	LINK	The film jammed in the movie *reel*, near the end.

during the 10 s between sentences. We then selected for the nonanomalous condition the sentences with the most accurately spelled words (median correct 100%). To discourage development of homophone search strategies and a possible bias towards judging sentences as anomalous, filler sentences came in a single *nonanomalous* version that contained no obvious homophones.

Procedure

The experiment involved two phases: *auditory exposure* and *test*. Ten practice trials preceded each phase.

Auditory exposure phase. Exposure trials began with a 500 ms "READY" signal centred on the monitor, followed by an auditory sentence over headphones, and a written prompt for participants to repeat the sentence word for word from memory as quickly as possible. Auditory exposure materials were the 20 fillers and nonanomalous versions of the 40 experimental sentences, recorded digitally in standard American dialect. Each participant heard half the fillers and half the experimental sentences in the auditory exposure phase (randomly ordered and counterbalanced across participants).

Test phase. Test phase instructions were presented by computer and verbally reiterated by the experimenter. Instructions informed participants that they would see sentences presented one word at a time over the fixation point, followed by two questions about each sentence. Instructions also indicated that some sentences would be anomalous in grammar or spelling, and examples of anomalies were given. A 500 ms "READY" signal began each trial, followed by a sentence presented at 90 ms per word using rapid serial visual presentation (RSVP; see e.g., MacKay, Miller, & Schuster, 1994, for procedural details). We employed RSVP in order to prevent rereading and assure that phonological effects were not due to "second pass" processing (see, e.g., Waters et al., 1992).

Immediately after each sentence, participants saw the comprehension-memory question, *Did you hear that exact sentence before?*, and pressed a "yes" button if the sentence had exactly the same meaning as an auditory sentence heard earlier, and a "no" button if even a single word changed the meaning of the sentence. Then came the visually presented anomaly detection question, *Was that a valid sentence?*, and participants pressed a "yes" button for sentences with completely correct grammar and spelling, and a "no" button otherwise. After this second response, the computer advanced to the next trial.

During the test phase participants saw the 20 fillers randomly intermixed with 40 experimental sentences, 10 experimental sentences in each exposure condition (counterbalanced across participants). Half of the experimental sentences (counterbalanced across participants) were in the exposed condition, and were identical to sentences heard during the auditory exposure phase, except for the substituted phonologically identical, phonologically similar, and control words. For exposed experimental sentences, the correct response to both test questions was "yes" for nonanomalous versions, and "no" for the anomalous phonologically identical, phonologically similar, and control versions.

Results and discussion

Before all data analyses we removed trials in the test phase that involved equipment failures (0.23% of all test trials), or errors in shadowing the critical words in the auditory exposure phase (2.34% of all test trials). After removing trials from one phase, we removed corresponding trials from the other phase, so that we only analysed trials that were included in both the exposure and test phases. As noted in the description of our materials, word frequency did not differ between the experimental conditions. To further rule out frequency-based explanations of the phonological Armstrong effect, we confirmed that word frequency of the nonanomalous words (see Appendix) did not correlate reliably with anomaly-detection or comprehension-memory measures for either the phonologically identical or phonologically similar conditions. Moreover, separate correlations for conditions with prior auditory exposure and without prior auditory exposure were also nonsignificant (all ps > .10).

As in Experiment 1, Friedman ANOVA tests were used to test for main effects across conditions, followed by Wilcoxon tests for planned comparisons between conditions. Results of analyses by subject and by item are reported for each comparison.

Anomaly detection results

Figure 4 shows our anomaly detection measure, the mean proportion of "no" responses to the question: *Was that a valid sentence?* The effect of word type was examined overall, and for each level of exposure (exposed and unexposed) separately.

A Friedman ANOVA revealed a main effect of word type, $p < .001$ by subject and by item, which reflected three main differences between word type conditions: One was successful stimulus construction, i.e., more "no" responses overall to anomalous (phonologically identical, phonologically similar, or control) than nonanomalous versions, smallest Wilcoxon $z = 4.26$, $p < .001$ by subject and $z = 4.92$ $p < .001$ by item. The second effect of word type was reduced anomaly detection for phonologically identical than control versions, Wilcoxon $z = 2.58$, $p < .05$ by subject and $z = 1.93$, $p = .05$ by item, which we call the "phonological Armstrong effect". Whereas shared orthography, semantics, and surname nodes can contribute

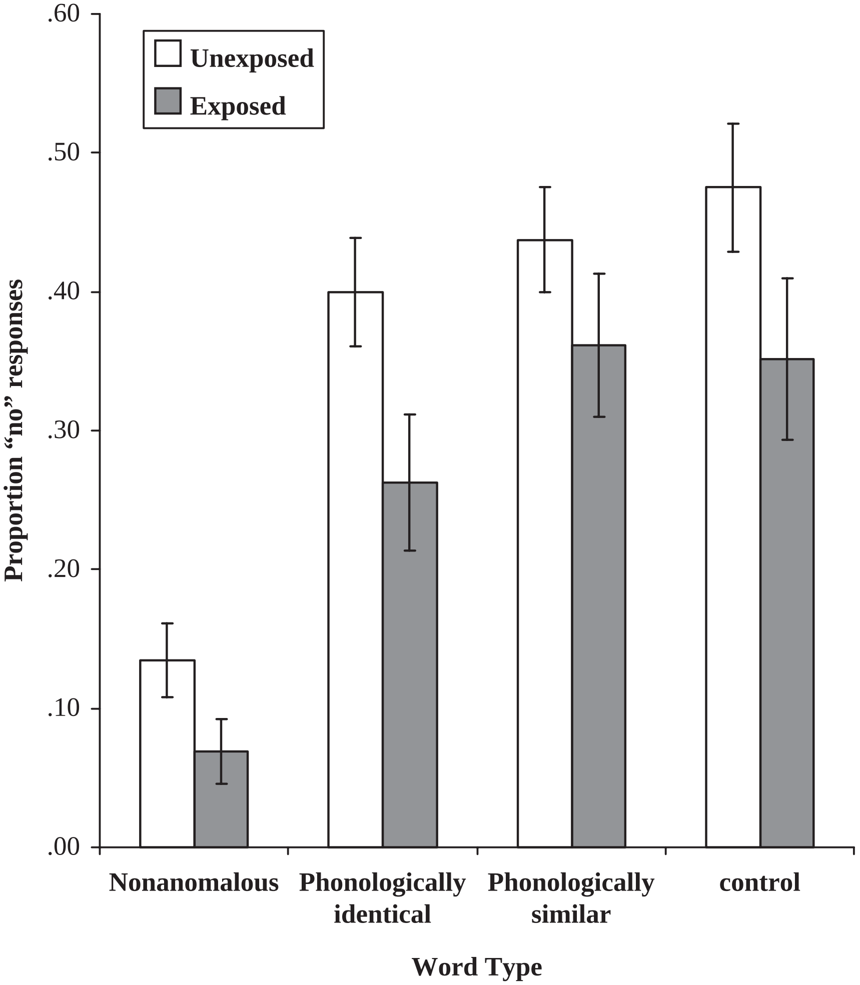

Figure 4. Mean proportion of "no" responses to the anomaly detection question *Was that a valid sentence?* by word type and exposure condition in Experiment 2. The error bars indicate ±1 SE.

to the standard Armstrong effect, only shared phonology could cause the phonological Armstrong effect in Experiment 2 (see the Methods section). As a new type of Armstrong effect involving word substitutions from classes other than proper names, the phonological Armstrong effect overcomes a limitation apparent in previous studies of Moses and Armstrong mistakes: the almost exclusive use of proper names as critical words.

The third finding of interest was a *nonsignificant* difference between the phonologically similar and control conditions, $p > .10$ both by subject and by item, indicating that the low-level (subsyllabic) phonological overlap with nonanomalous words had no more effect on anomaly detection than an unrelated (control) condition. This was further supported by marginally lower anomaly detection in the phonologically identical compared to phonologically similar condition, Wilcoxon $z = 1.82, p = .068$ by subject, $p > .10$ by item.

NST accurately predicted this nonmonotonic effect of word type. Under NST, participants often indicated that phonologically identical sentences were valid because they miscomprehended the substituted word (*mail*) as the nonanomalous word (*male*), the first-order basis for failing to detect anomaly under NST. However, anomaly detection was less common and not reliably different for control (*shed*) and phonologically similar (*mill*) substitutions under NST because the phonological syllable for neither *mill* nor *shed* links directly with the lexical node for *male*. As a consequence, neither *mill* nor *shed* transmits first-order priming to *male* so as to induce nonanomalous word miscomprehensions and phonological Armstrong effects. This finding is difficult to explain under a feature matching account, where the differential impact of direct (syllable) and indirect (subsyllabic) priming of a lexical representation is not taken into account.

Effects of exposure

Separate post hoc tests for exposed and unexposed sentences indicated that the nonmonotonic relation between anomaly detection and word condition was most apparent in exposed sentences (see Figure 4): There was no reliable difference between phonologically similar and control words, $z < 1$ by subject and by item, but there was less anomaly detection for phonologically identical than phonologically similar words, Wilcoxon $z = 2.02, p < .05$ by subject and $z = 1.40$, $p = .16$ by item. Although this same pattern was apparent for unexposed sentences, the difference between phonologically identical and phonologically similar words was not significant.

Present effects of exposure on anomaly detection supported NST predictions: less anomaly detection for exposed than unexposed sentences, with a nonmonotonic pattern involving no difference in anomaly detection between phonologically similar and control versions, and less anomaly detection for phonologically identical than phonologically similar versions, especially for exposed sentences. Under NST, initial exposure (shadowing) of the nonanomalous version strengthened both top-down semantic connections and bottom-up phonological connections to the lexical node for the nonanomalous word. During subsequent anomaly detection, these strengthened connections further enhanced the probability of activating the nonanomalous word in error, especially for anomalous sentences containing phonologically identical words, because phonologically identical words prime the lexical node for the nonanomalous word via the same bottom-up connections. For example, prior processing of the nonanomalous noun phrase *the dominant male* strengthened the top-down semantic connections to *male* and the bottom-up phonological connections from the phonological *male-mail* syllable to the lexical node for *male*. During visual presentation of the phonologically identical sentence, *In gorilla culture, the dominant mail must defend his status*, the lexical node for *male* therefore received exposure-enhanced priming via the strengthened bottom-up connections from the *male-mail* syllable to *male*. As a consequence, exposure-enhanced phonological priming selectively increased the probability of miscomprehending phonologically identical words as nonanomalous words. Because phonologically similar and control words do not share a syllable-to-lexical node connection with the nonanomalous words that prior exposure can strengthen, no similar exposure-enhanced bottom-up priming was possible in the phonologically similar and control conditions.

Present results did not support the response bias account of exposure effects in Kamas et al. (1996). Participants in Kamas et al. read correct statements (*Noah took two animals of each kind on the Ark*), half of which capitalised either the critical name (*NOAH took two animals of each kind on the Ark*) or the critical answer (*Noah took TWO animals of each kind on the Ark*). For the subsequent question-answering task, measures of sensitivity and bias indicated that "can't say" responses increased with capitalisation for both anomalous and nonanomalous *Moses* questions, suggesting that response bias explained why participants responded "can't say" more often for anomalous questions with capitalised rather than uncapitalised information during exposure. However, Kamas et al. did not include a condition with *no* prior exposure, ruling out a test of top-down effects predicted under NST, where exposing the nonanomalous version strengthens top-down connections that strongly prime the nonanomalous internal representation and thereby contribute to Moses mistakes.

Although the measures in Kamas et al. (1996) versus Experiment 2 differed, the Kamas et al. results do raise the question of whether response bias played a role in the present results. Although Experiment 2 was not designed to measure sensitivity and bias,[2] our separate post hoc tests for exposed versus unexposed

[2] We agree with Kamas et al. (1996) and Park and Reder (2004) that sensitivity and bias measures may be relevant to understanding Moses mistakes and related phenomena. However, with the design of Experiment 2 we could not legitimately compute sensitivity and bias by condition (see Stanislaw & Todorov, 1999) and recent criticisms (see Pastore, Crawley, Berens, & Skelly, 2003) call into question the validity of the nonparametric measures used in Kamas et al. For example, Pastore et al. (2003) question the common assumption that A' and B'' are independent and demonstrate that the types of designs that we and others use systematically underestimate A', an error that increases with higher sensitivities. Taken together, these criticisms indicate that nonparametric analyses may be particularly problematic for specific comparisons between sensitivity and bias measures across conditions.

sentences indicated that exposed more so than unexposed sentences exhibited word type effects. Such differential effects of exposure across word type are difficult to explain in terms of exposure-induced response bias. Additionally, exposure yielded a marginal decrease in "no" responses in the nonanomalous condition, Wilcoxon $z = 1.76$, $p = .079$ by subject, so that exposure increased *correct* responses rather than false alarms. This result is incompatible with a response bias explanation, but comports with the explanation of present exposure effects under NST, where exposure strengthens the connections for correct information, impeding comprehension of anomalous sentences and facilitating comprehension of nonanomalous sentences.

Why did evidence for response bias emerge in Kamas et al. (1996) but not in the present false alarm rates? Several procedural details distinguish Kamas et al. from the present study, e.g., use of proper names as critical words and indirect rather than direct tests of anomaly detection. However, the most likely explanation lies in the Kamas et al. "emphasis manipulation" because capitalisation of critical words in their subsequent experiments (3a and 3b) led to a similar response bias *without* prior exposure.

Comprehension-memory

Figure 5 shows our comprehension-memory measure, the mean proportion of "no" responses to the question: *Did you hear that exact sentence before?* As can be seen in Figure 5, accuracy was high for all word types in the unexposed condition, and a Friedman ANOVA revealed no word type effect, $p > .1$ both by subject and by item. Thus, subsequent analyses are reported for the exposed conditions only. A Friedman ANOVA on these data yielded a main effect of word type, $p < .001$ for both by subject and by item analyses.

Examining word type effects for the exposed versions, comprehension-memory results mirrored anomaly detection results and comport with NST: "No" responses were less common for phonologically identical than control versions, Wilcoxon $z = 2.90$, $p < .01$ by subject and $z = 3.50$, $p < .001$ by item, and for phonologically identical than phonologically similar versions, Wilcoxon $z = 2.66$, $p < .01$ by subject and $z = 2.10$, $p < .05$ by item, with no reliable difference between phonologically similar and control versions, $p > .1$ both by subject and by item.

GENERAL DISCUSSION

The present results strengthen the case that the Moses, Armstrong, and affiliated family of "illusions" involve the same processes as normal error-free sentence comprehension, e.g., priming from bottom-up and top-down sources, processes for integrating lexical and propositional representations,

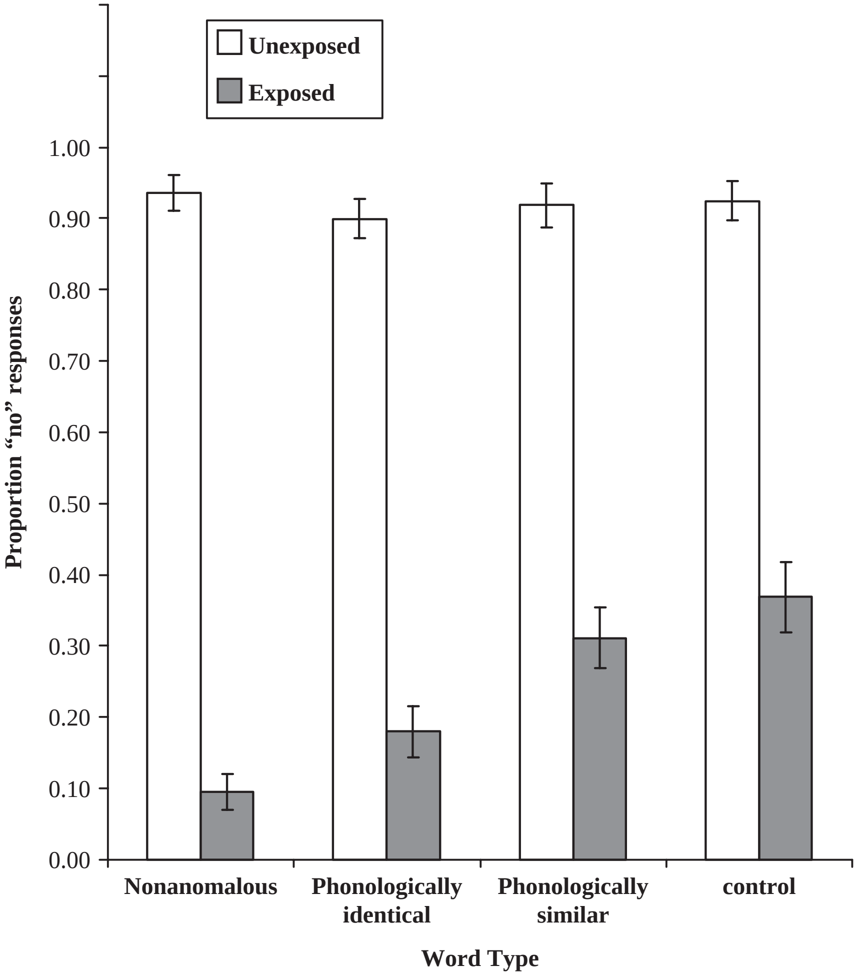

Figure 5. Mean proportion of "no" responses to the comprehension-memory question *Did you hear that exact sentence before?* by word type and exposure condition in Experiment 2. The error bars indicate ±1 SE.

and strengthening of specific connections due to recent exposure. Experiment 1 replicated the standard Armstrong effect and demonstrated new unknown Armstrong effects involving names with little or no associated propositional information in semantic memory. When unknown names of the same-gender were substituted for the expected name, unknown Armstrong effects exceeded the standard Armstrong effect in magnitude, and when unknown names with different-gender substituted the expected name, the unknown Armstrong effect was smaller but nevertheless reliable relative to the control condition. Together with corroborating data from the comprehension-memory task, these unknown Armstrong effects indicated that lexical- and proposition-level processes contributed autonomously to anomaly detection: *Lexical-level* processes precluded anomaly detection when participants *inaccurately* comprehended the substituted name as the expected name. However, *proposition-level* processes precluded anomaly detection when participants *accurately* comprehended the substituted name and could integrate this lexical comprehension into the sentence context without the comprehension difficulties that signal the presence of anomaly.

Experiment 2 demonstrated a phonological Armstrong effect for substituted content words that differed from the expected word in meaning and orthography but not phonology. Although prior exposure to the nonanomalous version of the sentence increased the difference between anomaly detection in phonologically identical and phonologically similar conditions, it did not lead to a significant difference between phonologically similar and control conditions. This nonmonotonic effect of phonological similarity, and the decreased anomaly detection following exposure to the nonanomalous version of a sentence, supported NST predictions.

Neither the unknown nor phonological Armstrong effects reflected inadequate knowledge or word-level misperception because data were excluded for participants who did not know the critical words or failed to repeat them correctly in the initial shadowing task. Nor were the unknown and phonological Armstrong effects attributable to the degree of semantic overlap between the expected and substituted words: Semantic overlap was greater for control than unknown names in our demonstration of the unknown Armstrong effect (Experiment 1), and was minimal and equated across conditions in our demonstration of the phonological Armstrong effect (Experiment 2). Nor was the phonological Armstrong effect attributable to orthographic overlap between the expected and substituted names: Orthographic overlap did not differ in degree for the phonologically identical and phonologically similar conditions in Experiment 2. Moreover, the phonological Armstrong effect was not related to the frequency of the expected word, a variable that may impact homophone error rates when processing single words in isolation (see, e.g., Ziegler, Van Orden, & Jacobs, 1997). Word frequency of the nonanomalous words (see Appendix) did not correlate reliably with "no" rates in the anomaly detection or comprehension-memory tasks, either for the homophone (phonologically identical) condition or for our partially overlapping (phonologically similar) condition. Finally, neither the unknown nor phonological Armstrong effects reflected momentary attentional lapses that caused people to "miss" or "misperceive" the critical words: Participants always attended to and correctly perceived the critical words in our data because we only scored answers to anomaly detection and comprehension-memory questions when the critical words were correctly shadowed and therefore attended to and correctly perceived.

Other theoretical frameworks

As discussed earlier, the main alternative to the NST "misunderstanding" account of Moses mistakes is a class of feature-matching models, represented by the Partial Match Hypothesis (PMH; Barton & Sanford, 1993; Kamas et al., 1996; Reder & Cleeremans, 1990; Reder & Kusbit, 1991; van Oostendorp & Kok, 1990). PMH successfully explains several aspects of Moses illusions, but many specific findings in the present study challenge PMH and other feature-matching models. An example is the reduced anomaly detection for unknown different-gender names versus control names in Experiment 1. PMH would predict *greater* anomaly detection for unknown different-gender names than control names, i.e., the reverse of the present findings. As discussed in the introduction, control names share gender and celebrity status information with the nonanomalous names, and this overlap should increase semantic cohesion and reduce anomaly detection for control names relative to unknown different-gender names, where this overlap is eliminated. The same holds for control names versus unknown same-gender names. PMH would predict *greater* anomaly detection for unknown same-gender names than control names (albeit not as much greater as for unknown different-gender names): Controls names but not unknown same-gender names have overlapping gender information that should contribute to semantic cohesion and reduce anomaly detection, i.e., the reverse of the present findings. The basic mechanisms of PMH therefore provide no way of explaining why less anomaly detection occurred for unknown names than for control names in Experiment 1.

The unknown and standard Armstrong effects also challenge a second assumption in the original PMH, that phonological features do not influence cohesion, and thus do not play a role in Moses mistakes. Contrary to this PMH assumption, Experiment 1 demonstrated reduced anomaly detection when the expected and substituted names overlapped equally in semantics but unequally in phonology in the unknown and Armstrong conditions. These results raise an important question: Can a simple extension of PMH principles to phonological levels explain the unknown and standard Armstrong effects (as per L. M. Reder, personal communication, 2 September 2005)? Under this PMH extension, participants initially construct a partial phonological analysis of the substituted words that render them phonologically indistinguishable from the expected words, thereby preventing anomaly detection. However, this PMH extension cannot readily explain an important aspect of available data: the accurate shadowing of the critical words that served as a precondition for analysing standard Armstrong effects in Shafto and MacKay (2000) and unknown Armstrong effects in Experiment 1. If participants initially constructed a *partial* phonological analysis that rendered the expected and substituted words phonologically indistinguishable during sentence comprehension, then how could they accurately shadow or produce the *full* phonology for the anomalous words prior to anomaly detection in these studies?

This is not to say that Experiment 1 results rule out *all* feature-matching models. Most of the results in Experiment 1 are compatible with a different type of feature-based model known as Match–Mismatch Theory (MMT). Like PMH, MMT addresses Moses and Armstrong mistakes and includes mechanisms whose sole function is anomaly detection. However, MMT is more general than PMH, e.g., applying to learning (e.g., Cormier, 1981) and visual perception (e.g., Bridgeman, 2003; Stark & Bridgeman, 1983; Teuber, 1960), and, unlike PMH, MMT readily accounts for the detection of orthographic, phonological, and semantic anomalies in comprehending isolated words (see e.g., MacKay, 1972).

Under MMT, anomaly detection mechanisms compute matches and mismatches between the phonological, orthographic, and semantic features of an input with internally generated expectations based on context or prior experience. Moses and Armstrong mistakes therefore occur because phonological, orthographic, or semantic matches between an expected and substituted word greatly outnumber mismatches, reducing the probability of anomaly detection. For example, because the concepts *Noah*, *the Ark*, and *animals of each kind* have co-occurred frequently in prior experience, listeners expect the word *Noah* in the question, *How many animals of each kind did* **** *take on the Ark?* The *Moses* substitution therefore often passes undetected under MMT because the expected *Noah* overlaps with *Moses* in phonology, orthography, and (especially) semantics, yielding a large number of feature matches that reduce the probability of anomaly detection.

MMT also readily explains the standard Armstrong effect, i.e., greater anomaly detection for control than Armstrong names. The standard Armstrong effect arises under MMT because anomaly detection mechanisms compute more matches and fewer mismatches between phonological and semantic features of the Armstrong name (*Gerald Ford*) and the expected name (*Henry Ford*) than between the control name (*Herbert Hoover*) and the expected name.

MMT can also readily accommodate unknown Armstrong effects (unlike PMH). The unknown same-gender condition leads to more mistakes than the Armstrong condition under MMT because unknown same-gender names (*Michael Armstrong*) share phonological (surname) and semantic (gender) information with the expected name (e.g., *Neil Armstrong*), but lack semantic features that *mismatch* the expected name. Unknown same-gender names therefore cause a greater than standard Armstrong effect (with *less* anomaly detection for same-gender Armstrong names than control names) because fewer semantic features *mismatched* with the expected names for same-gender than control names under MMT. Finally, MMT can account for the smaller unknown Armstrong effect in the different-gender than same-gender condition in Experiment 1 because different-gender names, by virtue of being the wrong gender, have more mismatching features and fewer matching features than same-gender names.

However, two effects in Experiment 2 are problematic for MMT. One is the effect of prior exposure. In its current form, MMT cannot predict our observed reduction in anomaly detection with prior exposure because the number of matching versus mismatching features do not differ before versus after exposure. However, one can imagine a modified version of MMT in which feature matches and mismatches are weighted to reflect exposure-linked salience before being tallied (see Levy, 1983). This version of MMT can predict an overall exposure-linked reduction in anomaly detection, with monotonic increases in anomaly detection across the phonologically identical, phonologically similar, and control conditions for both exposed and unexposed sentences in Experiment 2. The reason is that an exposure-linked decrease in the salience of mismatches with the nonanomalous word will decrease anomaly detection overall, without altering the relative numbers of matching versus mismatching features between the expected (nonanomalous) word and the various types of anomalous words in either exposed or unexposed sentences. However, for exposed sentences, Experiment 2 results indicated a nonmonotonic effect that is unpredicted under MMT: greater anomaly detection for phonologically similar than for phonologically identical versions, but no difference in anomaly detection between phonologically similar and control versions.

SUMMARY AND CONCLUSIONS

The unknown and phonological Armstrong effects are readily explained in theories with two general characteristics. One is a focus on the nature of how knowledge is structured and used, in this case the hierarchically organised representational units and processes underlying the comprehension, production, and memory for language. The second general characteristic is a focus on the primary goal of language processing, to successfully *integrate* linguistic meanings and stored knowledge into a coherent conceptual representation. Theories like NST share these characteristics, and explain anomaly detection, miscomprehension, and errors or illusions as side effects of the normal mechanisms for creating integrated conceptual representations. By contrast, feature-matching theories with specific applications to

Moses illusions and related effects (like PMH) or with broader applications (like MMT) do not share these characteristics and cannot readily explain the present results, even with *post hoc* modifications. Moreover, feature-matching accounts of anomaly detection call for special anomaly detection mechanisms that serve no other function. Although the brain has evolved special mechanisms to detect *novelty*, an additional evolution-based mechanism devoted to detecting sentence-level anomalies seems unlikely. The reason is that errors and anomalies are relatively rare and inconsequential for everyday language comprehension (see e.g., Brown & Hanlon, 1970; Pinker, 1999) and production (Fromkin, 1973), which suggests that evolutionary pressures for developing *special* mechanisms for detecting sentence-level anomalies are negligible.

In conclusion, we return to our original question: Why do people sometimes think they comprehend, but don't? The answer seems to lie not with the failure of mechanisms dedicated to anomaly detection but with the nature of general mechanisms for language comprehension and memory involving both error-free and anomalous information. What is needed in future modelling efforts and research on error detection is an approach that integrates comprehension and memory, such that experience can strengthen exactly those bottom-up (phonological) and top-down (semantic) processes involved in comprehension.

Under this integrative approach, the family of *memory* illusions (Schacter, Coyle, Fischbach, Mesulam, & Sullivan, 1995) and the present phonological and unknown Armstrong effects reflect fundamentally similar principles, despite the differing time factors, procedures, stimulus modalities, and descriptive labels in studies examining memory illusions versus Moses-like illusions. Studies of memory illusions (e.g., Loftus, Feldman, & Dashiell, 1995) typically involve the implantation of false memories: Participants see a visual sequence depicting a complex, ambiguous, and unfamiliar event, and then hear an account of the event containing factual errors. Participants then complete a set of distractor tasks followed by a comprehension-memory test of for the original event. The result is a high proportion of false memories based on the factual errors.

Bartlett and others obtained similar results in the original research on memory illusions (Bartlett, 1932; and Carmichael, Hogan, & Walter, 1932). To illustrate, on each trial in Carmichael et al. (1932), participants heard words, e.g., "eye-glasses", with conceptual links to some aspect of a subsequently presented visual form that was ambiguous, e.g., two circles connected by a short straight line that might also represent a dumbbell. During subsequent free recall of the visual stimuli, participants often misrepresented the visual forms based on the misleading words. As Bartlett (1932) noted in discussing this and other memory illusions, encoding and retrieving memory traces for novel, ambiguous, and anomalous or culturally foreign information requires an "effort to understand" based on preexisting memory structures involving linguistic and cultural knowledge.

Like the present data, Bartlett's (1932) evidence for "understanding processes" that "reconstruct" perceptual inputs and memories carries two implications: that the conceptual separation of memory from comprehension, language, and culture in general paints a misleading picture of the human mind; and that human information processing is designed not to *evaluate* coherence and detect anomaly via special anomaly detection mechanisms, but to *achieve* coherence by integrating all available and relevant information, a process with side effects that sometimes distort retrieval and encoding processes. This goal of achieving coherence likewise led to inaccurate recall in the present comprehension-memory tests, which reflected mistaken integration of anomalous information with error-free aspects of a sentence.

Original manuscript received August 2008
Revised manuscript received March 2009
First published online November 2009

REFERENCES

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, UK: Cambridge University Press.

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory and Cognition*, *21*, 477–487.

Bridgeman, B. (2003). Why every perceptual psychologist should know about eye movements. *American Journal of Psychology*, *116*, 315–318.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 11–54). New York: Wiley.

Brysbaert, M., Grondelaersb, S., & Ratinckxa, E. (2000). Sentence reading: Do we make use of orthographic cues in homophones? *Acta Psychologia*, *105*, 31–56.

Carmichael, L., Hogan, H. P., & Walter, A. A. (1932). An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of Experimental Psychology*, *15*, 73–86.

CMU Pronouncing Dictionary (version. 0.6) [Data file]. (1998). Carnegie Mellon University, Speech at CMU Web site. Retrieved from ftp://ftp.cs.cmu.edu/project/speech/dict/cmudict.0.6

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505. Updated version retrieved from http://www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm

Cormier, S. M. (1981). A match-mismatch theory of limbic system function. *Physiological Psychology*, *9*, 3–36.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*, 540–551.

Fromkin, V. A. (1973). *Speech errors as linguistic evidence*. The Hague, The Netherlands: Mouton.

Kamas, E. N., Reder, L. M., & Ayers, M. S. (1996). Partial matching in the Moses illusion: Response bias not sensitivity. *Memory and Cognition*, *24*, 687–699.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–146). New York: Wiley.

Levy, B. A. (1983). Proofreading familiar text: Constraints on visual processing. *Memory and Cognition*, *11*, 1–12.

Loftus, E. F., Feldman, J., & Dashiell, R. (1995). The reality of illusory memories. In D. L. Schacter, J. T. Coyle, G. D. Fishbach, M. M. Mesulam, & L. E. Sullivan (Eds.), *Memory distortion: How minds, brains and societies reconstruct the past* (pp. 47–68). Cambridge, MA: Harvard University Press.

MacKay, D. G. (1972). Input testing in the detection of misspellings. *American Journal of Psychology*, *85*, 121–127.

MacKay, D. G. (1973). Aspects of the theory of comprehension, memory, and attention. *Quarterly Journal of Experimental Psychology*, *25*, 22–40.

MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills.* New York: Springer-Verlag.

MacKay, D. G., & James, L. E. (2002). Aging, retrograde amnesia, and the binding problem for phonology and orthography: A longitudinal study of "hippocampal amnesic" HM. *Aging, Neuropsychology, and Cognition*, *9*, 298–333.

MacKay, D. G., Miller, M. D., & Schuster, S. P. (1994). Repetition blindness and aging: Evidence for a binding deficit involving a single theoretically-specified connection. *Psychology and Aging*, *9*, 251–258.

Park, H., & Reder, L. M. (2004). Moses illusion: Implication for human cognition. In R. F. Pohl (Ed.), *Cognitive illusions* (pp. 275–291). Hove, UK: Psychology Press.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin and Review*, *10*, 556–569.

Pinker, S. (1999). *Words and rules: The ingredients of language.* New York: Basic Books.

Reder, L. M., & Cleeremans, A. (1990). The role of partial matches in comprehension: The Moses illusion revisited. In A. Graesser & G. Bower (Eds.), *The psychology of learning and motivation* (Vol. 25, pp. 233–258). New York: Academic Press.

Reder, L. M., & Kusbit, G. W. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *Journal of Memory and Language*, *30*, 385–406.

Schacter, D. L., Coyle, J. T., Fishbach, G. D., Mesulam, M. M., & Sullivan, L. E. (Eds.). (1995). *Memory distortion: How minds, brains and societies reconstruct the past*. Cambridge, MA: Harvard University Press.

Shafto, M. A., & MacKay, D. G. (2000). The Moses, Mega-Moses, and Armstrong illusions: Integrating language comprehension and semantic memory. *Psychological Science*, *11*, 372–378.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments and Computers*, *31*, 137–149.

Stark, L., & Bridgeman, B. (1983). Role of corollary discharge in space constancy. *Perception and Psychophysics*, *34*, 371–380.

Teuber, H.-L. (1960). Perception. In J. Field, H. W. Magain, & V. E. Hall (Eds.), *Handbook of physiology* (Vol. 3, pp. 1595–1688). Washington, DC: American Physiological Society.

van Oostendorp, H., & Kok, I. (1990). Failing to notice errors in sentences. *Language and Cognitive Processes*, *5*, 105–113.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound and reading. *Memory and Cognition*, *15*, 181–198.

Waters, G. S., Caplan, D., & Leonard, C. (1992). The role of phonology in reading comprehension: Implications of the effects of homophones on processing sentences with referentially dependent categories. *Quarterly Journal of Experimental Psychology*, *44A*, 343–372.

Ziegler, J. C., van Orden, G. C., & Jacobs, A. M. (1997). Phonology can help or hurt the perception of print. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 845–860.

APPENDIX

Experimental target words used in Experiment 2

Nonanomalous			*Phonologically identical*			*Phonologically similar*			*Control*		
Word	*WF*	*"no" rate*	*Word*	*WF*	*"no" rate*	*Word*	*WF*	*"no" rate*	*Word*	*WF*	*"no" rate*
BAIL	7	0.12	BALE	5	0.25	BOIL	12	0.62	TANK	12	0.62
BEACH	61	0.00	BEECH	6	0.37	BATCH	5	0.37	CRANE	5	0.62
BEAR	57	0.50	BARE	29	0.86	BLUR	3	0.50	DOOM	3	0.62
BOAR	1	0.25	BORE	24	0.71	BEER	34	0.75	CORN	34	0.62
BORED		0.00	BOARD	239	0.12	BEARD	26	0.37	CHARM	26	0.50
BUY	70	0.12	BYE	2	0.75	BAY	57	0.37	LEG	58	0.71
DEER	13	0.12	DEAR	54	0.25	DOUR	2	0.75	LAME	2	0.75
DUAL	9	0.12	DUEL	5	0.12	DIAL	1	0.00	SMOG	1	0.25
DUE	142	0.00	DEW	3	0.62	DIE	73	0.00	FIG	72	0.71
FAIR	77	0.00	FARE	7	0.25	FEAR	127	0.37	NOTE	127	0.25
FOUL	4	0.00	FOWL	1	0.25	FAIL	37	0.14	BONE	33	0.37
GROWN	43	0.00	GROAN	1	0.50	GRAIN	27	0.43	RANCH	27	0.25
HAIL	10	0.25	HALE	2	0.25	HULL	13	0.37	GRIN	13	0.37
HAIR	148	0.12	HARE	1	0.25	HIRE	15	0.62	BURN	15	0.50
HALL	152	0.00	HAUL	5	0.37	HILL	72	0.00	BOAT	72	0.37
HEAL	2	0.25	HEEL	9	0.37	HOWL	4	0.37	VEST	4	0.25
LAIN	4	0.00	LANE	30	0.29	LAWN	15	1.00	ROPE	15	0.40
LOAN	46	0.00	LONE	8	0.25	LEAN	20	0.25	BOLD	21	0.62
MALE	37	0.00	MAIL	47	0.25	MILL	11	0.25	SHED	11	0.00
NAVAL	33	0.00	NAVEL	2	0.00	NOVEL	59	0.12	MINOR	58	0.25
PAIR	50	0.12	PARE	2	0.29	PAIN	88	0.62	WAIT	94	0.50
PALE	58	0.25	PAIL	4	0.50	PILL	15	0.50	CLUE	15	0.33
PEEL	3	0.12	PEAL	1	0.50	PULL	51	0.40	REAR	51	0.50
PIER	3	0.00	PEER	8	0.37	PYRE	1	0.75	WAND	1	0.50
POLE	18	0.25	POLL	9	0.12	PILE	25	0.50	HARM	25	0.25

(*Continued*)

Appendix (*Continued*)

Nonanomalous			*Phonologically identical*			*Phonologically similar*			*Control*		
Word	*WF*	*"no" rate*	*Word*	*WF*	*"no" rate*	*Word*	*WF*	*"no" rate*	*Word*	*WF*	*"no" rate*
POUR	9	0.00	POOR	113	0.25	PURE	56	0.12	WILD	56	0.50
REEL	2	0.29	REAL	260	0.62	RAIL	16	0.37	LINK	16	0.37
ROLL	35	0.00	ROLE	104	0.37	RULE	73	0.50	GAIN	74	0.00
RUDE	6	0.00	RUED		0.12	RIDE	49	0.00	MILK	49	0.12
SAIL	12	0.17	SALE	44	0.12	SEAL	17	0.25	KISS	17	0.00
SOUL	47	0.12	SOLE	18	0.25	SELL	41	0.00	DIVE	23	0.50
STEAK	10	0.50	STAKE	20	0.37	STACK	9	0.37	GLOVE	9	0.25
STEEL	45	0.12	STEAL	5	0.12	STALL	18	0.37	TWIST	18	0.50
SUN	112	0.00	SON	166	0.43	SIN	53	0.50	WIN	55	0.50
SWEET	70	0.00	SUITE	27	0.62	SWEAT	23	0.50	CRAFT	23	0.25
TALE	21	0.12	TAIL	24	0.37	TILE	16	0.37	GOWN	16	0.62
TEE	5	0.00	TEA	28	0.00	TOY	4	0.37	HEM	4	0.12
TIED		0.00	TIDE	11	0.37	TOAD	4	0.62	DIME	4	0.87
WEAK	32	0.00	WEEK	275	0.37	WALK	100	0.50	CLAY	100	0.50
WEAR	36	0.25	WARE	1	0.29	WIRE	42	1.00	EASE	42	0.37

Word =critical experimental target word, listed by condition; WF =word frequency; "no" rate =proportion responding "no" per item to question "Was that a valid sentence?"