Available online at www.sciencedirect.com
ScienceDirect
Journal of Experimental Social Psychology 43 (2007) 287–294
Journal of Experimental Social Psychology
www.elsevier.com/locate/jesp

Auditory Stroop reveals implicit gender associations in adults and children ☆

Steven B. Most [a,*], Anne Verbeck Sorber [b,*,1], Joseph G. Cunningham [b]

[a] Department of Psychology, Yale University, P.O. Box 208205, New Haven, CT 06520-8205, USA
[b] Psychology Department, Brandeis University, MS 062 P.O. Box 549110, Waltham, MA 02454, USA

Received 12 June 2005; revised 28 December 2005
Available online 17 April 2006

Abstract

Gender provides a powerful social heuristic for structuring incoming information. Thus, it may be difficult to attend to aspects of a person's sex without also activating irrelevant gender associations. In two experiments, an auditory Stroop revealed implicit gender associations. Participants categorized the sex of voices saying names and words stereotypically associated with male, female or neutral gender roles. Both adults and children were slower when the voice's sex was stereotypically incongruent with the spoken word or name. Although both groups showed such interference, children—who are generally less flexible about gender roles—showed more interference in response to gender-stereotypical words (e.g., football) than names (e.g., Rachel), whereas adults showed the opposite pattern. Given the simplicity of this task, the auditory Stroop might be used both to tap into implicit gender associations and to investigate their development.

Keywords: Gender stereotypes; Auditory Stroop; Implicit stereotypes; Gender; Sex roles; Gender schemas; Gender categorization; Automatic stereotypes

Through our daily interactions with the world, we organize incoming information and integrate it—through no conscious act of will—into clusters. In many cases this is advantageous; we are bombarded by so much information at any given moment, that attending to each feature independently would leave us encumbered. Many tasks, however, require that we attend only to particular features of a stimulus, and it is here that our tendency to process items "holistically"—i.e., as a whole—can get in our way. No task better illustrates the difficulty of attending to one feature of a stimulus while ignoring another than the Stroop task, in which people have difficulty rapidly classifying the colors of letter-strings that spell out the names of conflicting colors (e.g., the word "red" printed in blue ink; Stroop, 1935). Through experience, learned associates can cause such interference as well; for example, words closely related to color-names (e.g., "sky") cause interference when printed in an incongruent color, albeit to a lesser degree than incongruent color-names themselves do (Klein, 1964; Scheibe, Shaver, & Carrier, 1967). The role of practice and experience in determining the structure of incoming information has been underscored through other paradigms. For example, when people had to judge whether the top halves of cars were the same or different as each other while ignoring

☆ The first experiment was supported in part by a Children's Studies Mentoring Award to the first author by Children's Studies at Harvard. The second experiment was conducted as a part of the second author's dissertation. A.V. Sorber is now at Cyber Seas, Inc. Thanks to Jerome Kagan, Mark McManis, Richard J. McNally, Jason P. Mitchell, Nancy Snidman, Sei Jin Ko, David M. Marx, Marvin Chun, and W. Tecumseh Fitch for helpful discussions and constructive feedback regarding Experiment 1, and Malcolm Watson and Leslie Zebrowitz for feedback regarding Experiment 2. Thanks also to parents, children, and school programs that graciously agreed to participate in the children's experiment. Finally, thanks to Laura Carroll, Jeanne Fama, Julie Kalos, Laurie Santos, Susan Townsend, and Amy Wiseman for providing their vocal talents.
* Corresponding authors.
E-mail addresses: most@post.harvard.edu (S.B. Most), anne@sorber.us. (A.V. Sorber), cunningh@brandeis.edu (J.G. Cunningham).
[1] Present address: Cyber Seas, Inc. 5 Dolloff Farm Drive, Exeter, NH 03833, USA.

0022-1031/$ - see front matter
doi:10.1016/j.jesp.2006.02.002

the bottom halves of the cars, which could vary independently, car experts had more difficulty ignoring the bottom halves than car novices did, and they thus experienced more interference from them (Gauthier, Curran, Curby, & Collins, 2003). In other words, their very expertise made them more likely to automatically integrate task-irrelevant, along with task-relevant, information.

The role of learning and pre-existing knowledge in determining how people structure information is important to understand. Furthermore, one need not seek out special populations, such as car experts, in order to see the organizing effects of pre-existing knowledge. People from an early age become experts at discerning social cues and categories. Thus, social categories might constitute an important, and perhaps automatic, vehicle through which people structure information in everyday life.

Of all social categories, those based on gender are possibly the most salient. All human cultures assign some adult roles according to gender and expect their children to be socialized to these roles (Bem, 1981; Glick & Fiske, 1999). Given the early age at which children begin to categorize people as male or female (Leinbach & Fagot, 1993), play with sex-typed toys (e.g., Bradbard, Martin, Endsley, & Halverson, 1986; O'Brien & Huston, 1985), segregate themselves with same-sex peers (Maccoby, 1988; Martin, 1999), and make sex-based inferences about appropriate behavior (Biernat, 1991; Kuhn, Nash, & Brucken, 1978), the implementation of gender-based associations might eventually become effortless, fast, implicit, and difficult to inhibit (e.g., Banaji & Hardin, 1996), characteristics often regarded as hallmarks of *automaticity* (Logan, 1988). Notably, such notions are consistent with the central claim of *gender schema* theory: once children come to recognize their own sex and the sex of people around them, they become motivated to seek out and actively organize incoming information according to gender-appropriateness (Bem, 1981; Martin & Halverson, 1981). Research stemming from this perspective has investigated the degree to which the resultant knowledge structures (or schemas) influence how individuals think, perceive, and behave towards others, the degree to which they subsequently influence the further organization of knowledge, and the degree to which individual differences can be predicted by the strength of such schemas (for reviews, see Martin, 1999; Martin & Ruble, 2004).

To date, most assessments of gender schema content have relied on explicit measures (e.g., Edelbrock & Sugawara, 1978; Liben & Bigler, 2002; Williams, Bennett, & Best, 1975). For example, questionnaires sometimes ask participants to indicate whether and to what extent they associate various traits with males and females (e.g., Williams et al., 1975). In other cases, they are shown objects and are asked how exclusively they are used by males versus females (e.g., Edelbrock & Sugawara, 1978). Although such studies have been helpful in revealing explicit gender knowledge, the very fact that they rely on individuals to explicitly recognize their own associative biases prevents them from tapping into a full spectrum of stereotypic associations. It is possible that people wish not to express or are truly not aware of all their gender associations. Indeed, calls have been made for broader, more complex arrays of measures of stereotypes (Martin, 1999).

To complement explicit questioning, one might utilize implicit measures to assess attitudes and associations. For example, the Implicit Association Test (IAT) measures response times in order to assess the degree to which people implicitly evaluate the relative favorability of words associated with social categories like race, age, or gender (e.g., Greenwald, McGhee, & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2005). Implicit measures do not always correlate with explicit ones, and in some cases implicit measures might more successfully predict a person's behavior than explicit measures do (Dovido, Kawakami, Johnson, Johnson, & Howard, 1997; McConnell & Leibold, 2001). Furthermore, implicit and explicit patterns of responses may follow different developmental trajectories. For example, in adapting the IAT for use with children, it was found both that children as young as 6 years old showed implicit evidence of race-based evaluations and that implicit and explicit attitudes about race began to diverge by 10 years of age (Baron & Banaji, 2006). Notably, although alterations to the IAT have been shown to increase its child-friendliness (e.g., presenting words through speakers to surmount varying levels of reading proficiency; Baron & Banaji, 2006), certain core aspects of the task potentially limit its ability to reveal contents of early implicit schemas. For example, the task requires that participants maintain two sets of representations simultaneously (e.g., press the left button if the stimulus is a white face OR a bad word), making it potentially too unwieldy for younger children. In addition, because the task measures the ease with which different concepts and evaluations can be mapped onto each other, it is optimal for assessing people's attitudes towards particular concepts instead of the contents of those concepts per se.

Here, in two experiments, we developed and tested a simple and direct tool for indexing implicit gender schema contents. We evaluated the potential of an auditory variant of the Stroop task to reveal whether adults and children implicitly and automatically activate gender-associations when making simple judgments about a speaker's sex. Different types of stimuli likely vary in their centrality to a person's gender-schema, and to gauge this we used two general classes of stimuli: gender-stereotypical words (e.g., football) and gender-typical names (e.g., Rachel). Auditory Stroop effects have been found before (Cohen & Martin, 1975; Hamers & Lambert, 1972; Jerger, Martin, & Pirozzolo, 1988; Morgan & Brandt, 1989) even with children as young as three years old (Jerger et al., 1988). Some of these experiments have used judgments of speaker sex (e.g., Green & Barber, 1981, 1983), but in these cases stimuli have generally been male and female voices saying words like "man" and "girl," which doubled as actual descriptors of the voices themselves. When the sex of the voice was

Table 1
Words and names used in the auditory stroop with the adult sample

Female stereotyped words	Male stereotyped words	Gender neutral words	Female names	Male names
Bracelet	Baseball	Apple	Amy	Brian
Cheerleader	Captain	Door	Cindy	David
Doll	Football	Draw	Jenny	George
Lipstick	Gun	Paper	Jill	Henry
Lovely	Pirate	Pencil	Julie	Jason
Makeup	Punch	Spoon	Laurie	John
Nurse	Rough	Table	Nancy	Matthew
Perfume	Soldier	Taste	Rachel	Michael
Pink	Tackle	Walk	Sarah	Peter
Pretty	Tough	Window	Susan	Robert

incongruent with the descriptor, participants took longer to classify the sex of the voice. In one case, similar results were obtained when participants classified voices as belonging to "Joan" or "Dave" (Green & Barber, 1981, Experiment 4), but to our knowledge no experiments have tested for Stroop-like interference caused by stereotypical *associates*—such as those that might form a gender schema.

Experiment 1: Implicit gender associations in adults

Methods

Participants and design

Participants were 21 male and 21 female college-age students. One participant of each sex was dropped prior to analysis because of complications during testing: one had difficulty distinguishing between the different sex voices, and one—due to the sensitivity of the voice trigger—was dropped because of loud breathing. All participants claimed English as their first language. All auditory stimuli were presented within-subjects.

Stimuli

The stimuli were auditory presentations of 10 stereotypically masculine words, 10 stereotypically feminine words, 10 stereotypically gender-neutral words, 10 male names, and 10 female names (see Table 1).[2] All stimuli were presented through headphones by an IBM-compatible microcomputer running SuperLab software (Cedrus, 1996) and had been recorded from each of three different women, who read words and names into a computer microphone. Pseudo-male voices were created by using a computer-based sound editing program (Syntrillium, 1996) to lower the pitch of each sound clip while holding tempo constant. Thus, while the pseudo-male and female sound-clips differed from each other in pitch, their speaking rates, volumes, tones, and onset times were identical. While this procedure lowered the fundamental frequencies of the sound-clips, it did not alter the formant frequencies, which are an additional feature distinguishing between real male and female voices (Klatt & Klatt, 1990). To ensure that this manipulation was effective and that participants experienced the manipulated voices as being male, 10 additional participants listened to the same 3 female and 3 pseudo-male voices saying only the neutral words, and they made speeded key-press responses indicating the intended sex of speaker. Accuracy rate was above 99%, and there were no significant response time differences in response to female voices (mean = 678 ms, $SD = 171$ ms) versus pseudo-male voices (mean = 665 ms, $SD = 102$ ms), $t(9) = .47$, $p = .65$. Therefore, people seemed to have no trouble regarding the female voices as female and the manipulated voices as male.

For the purposes of actual experimental analyses, *congruent* word-voice pairs were those in which a female voice said a stereotypically feminine word or name and a pseudo-male voice said a stereotypically masculine word or name. *Incongruent* word-voice pairs were those in which a female voice said a stereotypically masculine word or name and a pseudo-male voice said a stereotypically feminine word or name.

Procedure

Participants first listened to samples of all the pseudo-male and female voices (saying the word "hello") and engaged in a trial run to ensure that they could differentiate between the two groups of voices. They then began the experimental trials.

Participants were instructed to classify the sex of each voice as quickly as possible by saying "boy" or "girl" into a microphone. The experimenter sat several feet to the side of the participant and noted trials where the voice-key was tripped accidentally and where classification mistakes occurred. These were removed before response time analyses.

Each word was presented by each of the three pseudo-male and three female voices, yielding a set of 300 stimulus presentations. The onset of each word was separated from the next by 3 s. There were two random orders for the presentation of the words, neither of which contained more than four successive presentations by the same sex voice.

[2] Stereotyped and neutral words were selected by having 20 volunteers (12 females, 8 males) sit at a computer keyboard, listen to an auditory presentation of 110 words (all words were said by a single voice), and indicate their male or female gender-association for each word as quickly as possible. Two additional neutral words were added on the basis of questionnaire ratings.

One order was the reverse of the other, and each was presented to half the participants. The orders were broken into three 5-min blocks of 100 words-presentations, and each block began with six practice word-presentations.

Results and discussion

Data coding

The variable of interest for all participants was the response time to classify the sex of the speakers. Type of Word was included as a variable in order to assess whether gender-specific names might show a bigger effect of congruence than stereotyped words. Response times were measured from the onset of each stimulus presentation. Those that were less than 400 ms were regarded as artifacts and were eliminated. Due to the frequency of stimulus presentation, no recorded response times were greater than 3000 ms. Response times greater or less than three standard deviations from each participant's means for congruent, incongruent, and neutral words and names were removed, as were trials where participants made incorrect responses or activated the voice-key accidentally. Classification errors were not distinguished from accidental voice-key activation during testing, so pure classification errors were not available as an additional source of data. In all, 6.33% of the word-trials were removed, leaving a set of 11,240 good data-points across participants: 2266 congruent stereotyped words (5.58% removed), 2241 incongruent stereotyped words (6.63% removed), 2246 congruent names (6.42% removed), 2216 incongruent names (7.67% removed), and 2271 neutral words (5.38% removed). Notably, whereas the stimuli in the congruent and incongruent conditions were identical except for pitch, the neutral stimuli comprised an entirely different set of words and thus were not ideal controls. Therefore, they were not included in the main analyses and were used only for rough comparison purposes afterwards. Except where noted, participants served as the main unit of analysis.

Analyses

A 2 (Participant Sex) × 2 (Congruence) × 2 (Type of Word) mixed measures ANOVA conducted on response times revealed a main effect of congruence, $F(1,38) = 80.94$, $p < .001$. Response times to incongruent word-voice pairs were consistently longer than those to congruent word-voice pairs. This main effect was significant both for stereotyped words, $t(39) = 4.99$, $p < .001$, and for names, $t(39) = 7.59$, $p < .001$ (see Table 2). There was also an interaction between Congruence and Type of Word, $F(1,38) = 9.98$, $p = .003$, indicating that the effect of congruence was greater for names than for stereotyped words. There were no main effects or interactions involving Sex of Participant; males and females showed the same patterns of responses.

To test whether the main effect and interaction generalized beyond the specific stimuli selected for the experiment, we conducted an additional analysis of variance using words as the main unit of analysis (see Clark, 1973), with Congruence as a within-word variable and Type of Word as a between-words variable. This analysis provided further support for both the main effect of Congruence, $F(1,38) = 72.85$, $p < .001$, and the interaction between Congruence and Type of Word, $F(1,38) = 6.68$, $p = .014$.

Table 2
Mean response times (ms) for gender-congruent, gender-incongruent, and neutral words and names in Experiments 1 and 2 (standard deviations are in brackets)

	Congruent	Incongruent	Neutral	Congruence effect (Incong-Cong)
Experiment 1: Adults				
Words	780	802	787	23
	(190)	(195)	(198)	(29)
Names	782	829		47
	(187)	(214)		(39)
Experiment 2: Children				
Words	841	907	846	66
	(159)	(196)	(162)	(90)
Names	873	902		28
	(165)	(193)		(77)

Note. Direct RT comparisons are valid only where the durations of the word presentations are equivalent, as differences in stimulus utterance-length might affect RT. As noted in the Experiment 1 methods section, such durations are balanced across congruence for words and names separately (but not neutral words). See footnote 5 for adult RTs to only words that were also used with children.

The effect of Congruence for words and names could be attributable either to facilitation in the congruent condition, interference in the incongruent condition, or both. Although the words in the neutral set of stimuli were not identical to those in the sets of stereotyped stimuli and were therefore not ideal controls, they provided a rough gauge for assessing the roles of facilitation and interference. Compared to their response times to neutral words, participants' response times in the congruent conditions were faster for stereotyped words, $t(39) = 2.09$, $p = .043$, though not significantly for names, $t(39) = .77$, $p = .443$ (see Table 2). Meanwhile, their response times in the incongruent conditions were slower for both stereotyped words, $t(39) = 3.42$, $p = .001$, and names, $t(39) = 8.10$, $p < .001$. Thus, the role of facilitation in the congruent conditions was less consistent than the role of interference in the incongruent conditions.

Faster response times in the congruent conditions did not occur at the expense of accuracy; in fact, the conditions that yielded slower response times also yielded decreased accuracy. Because classification errors were not distinguished from other participant errors during testing (e.g., accidental tripping of voice-key), pure classification errors were not available as a source of data. However, when all types of errors were pooled for each participant, a main effect emerged for Congruence, $F(1,38) = 22.36$, $p < .001$, as did an interaction between Congruence and Type of Word, $F(1,38) = 4.90$, $p = .033$. The mean number of errors per participant were as follows (out of a possible 60 per condition): M errors for congruent words = .25, $SD = .54$; M errors for incongruent words = .60, $SD = .78$; M errors for

Table 3
Words and names used in the auditory Stroop with the child sample

Female stereotyped words	Male stereotyped words	Gender neutral words	Female names	Male names
Cheerleader	Baseball	Door	Cindy	David
Lipstick	Football	Draw	Jenny	Henry
Makeup	Rough	Paper	Jill	John
Pretty	Soldier	Pencil	Nancy	Michael
Pink	Tough	Spoon	Rachel	Peter
		Table		
		Window		
		Apple		
		Taste		

congruent names = .38, $SD = .70$; M errors for incongruent names = 1.23, $SD = 1.27$; M errors for neutral words = .45, $SD = .78$.[3] When words were used as the unit of analysis, a main effect of Congruence again emerged, $F(1, 38) = 19.47$, $p < .001$, although the interaction between Congruence and Type of Word was weak, $F(1, 38) = 3.38$, $p = .074$.

In summary, when people classified speakers' voices as belonging to a male or a female, they were slower to do so when the speaker said a word stereotypically associated with the opposite sex. This suggests that even though people tried to attend to mere vocal qualities, they could not avoid processing the meanings of the spoken words and associating them with gender stereotypes so quickly as to interfere with the simple perceptual classification. When processing information about sex-based vocal qualities, it seems, the processing of gender-stereotyped information occurred as well. Although this effect occurred for both names and gender-stereotyped words, it was substantially larger for names. One possibility is that the more strongly a word is associated with one sex or the other, the more of a Stroop-like effect it will yield.

Experiment 2: Implicit gender associations in children

Because gender-associations are learned over time, children might not show the same Stroop-like effect that adults do, particularly with words that pertain to learned gender roles. On the other hand, at an explicit level, even young children have substantial knowledge of sex-role stereotypes (Kuhn et al., 1978), and those in middle childhood have been found to be less flexible regarding gender roles than older children and adults are (see Serbin, Powlishta, & Gulko, 1993, for a review). For example, in a study that recruited participants who were 8–17 years of age, the youngest group was found to be the least flexible (Katz & Ksansnak, 1994). It is possible that—consistent with gender schema theory—children will activate gender stereotype knowledge when simply judging a speaker's sex, perhaps at least as strongly as adults do. If so, the auditory Stroop might prove fruitful for expanding our understanding of gender stereotype development beyond that provided by explicit measures. In Experiment 2, we used the auditory Stroop to assess the automatic activation of gender stereotypes in a sample of 8- and 9-year old children.

Methods

Participants

Fifty-five third-graders were recruited from two elementary schools. Seven participants who did not complete the experiment were excluded from the analyses. The remaining 48 participants (23 females, 25 males) ranged in age from 97 to 118 months (mean age = 8 years, 11 months).

Stimuli

The auditory Stroop task was presented on a Compaq Presario 1650 laptop computer running SuperLab software (Cedrus, 1996), with stimuli drawn from the larger list of words used in Experiment 1 (see Table 3). Two random orders of 104 trials were generated, one the reverse of the other, and children responded to stimuli in one of these two orders. Each word was presented by two of the female voices from Experiment 1 and their pseudo-male counterparts. The onset of each word was separated from the next by 3 s.

Procedure

All children were tested individually in a room separate from the main classroom, facing the computer and with the female experimenter sitting beside them. Participants were asked to indicate whether each word was spoken by a male or female voice. Instead of indicating their answers verbally into a microphone, children used a keyboard to indicate their response, ensuring that RTs reflected true responses instead of the accidental tripping of a microphone.[4] Half the participants were told to press the "a" key for girl and "'" key for boy, while the other half were told the reverse, and the computer recorded

[3] Neutral words were not included in the analyses of variance.

[4] Piloting for the adults in Experiment 1 had suggested that key-presses were not optimal responses for that population, as adults tended to respond before the stimulus presentation had ended, thus diminishing effects stemming from the meanings of the words. Vocal responses took more time to generate and allowed the meanings of the stimuli to be processed. Children's key-presses, on the other hand, typically were slow enough to allow for semantic processing of the stimuli prior to response.

accuracy and RT. To help children keep track of the appropriate keys, the keys were marked with star-shaped stickers and schematic faces of a boy and girl were taped on the appropriate side of the computer screen. Children were told to answer as fast as they could, and they began with a practice session consisting of 10 female and pseudo-male voices saying "hello". They then continued with the actual experiment.

Data coding

Response times from correct trials were averaged for congruent and incongruent presentations of female-stereotyped words, male-stereotyped words, female names, male names, and gender-neutral words. As in Experiment 1, gender-neutral words were not included in the main analysis and were used only for rough comparison purposes afterwards. Except where noted, participants served as the main unit of analysis.

Results and discussion

A 2 (Participant Sex) × 2 (Congruence) × 2 (Type of Word) ANOVA revealed a main effect of Congruence, with longer RTs to incongruent than congruent trials, $F(1,46) = 24.49$, $p < .001$ (see Table 2). This main effect was significant both for stereotyped words, $t(47) = 5.08$, $p < .001$, and for names, $t(47) = 2.56$, $p < .013$. The main effect of Type of Word was weak, $F(1,46) = 3.14$, $p < .083$, and there was no main effect of Participant Sex. There was a significant interaction between Congruence and Type of Word, suggesting that—in contrast to Experiment 1—the effect of congruence was stronger for stereotyped words than for stereotyped names, $F(1,46) = 6.25$, $p = .016$ (see Table 2).

An additional analysis of variance using words as the main unit of analysis, with Congruence as a within-word variable and Type of Word as a between-words variable, again demonstrated a significant main effect of Congruence, $F(1,38) = 13.13$, $p = .001$. However, there was no main effect or interaction involving Type of Word. Thus, the effect of Congruence, though not necessarily that of Type of Word, appear to generalize beyond the specific stimuli chosen for this experiment.

Compared to their response times to neutral words, participants' response times were not faster for congruent stereotyped words, $t(47) = .43$, $p = .671$, and they were slower for congruent names, $t(47) = 2.52$, $p = .015$ (names elicited slower RTs than neutral words in general, regardless of congruence). Meanwhile, their response times in the incongruent conditions were slower for both stereotyped words, $t(47) = 4.50$, $p = .001$, and names, $t(47) = 4.67$, $p < .001$ (see Table 2). As with the adults, this suggests a smaller role for facilitation in the congruent conditions than for interference in the incongruent conditions, although—as evidenced by the overall slower responses to names than neutral words—neutral words themselves did not serve as optimal control stimuli.

For error rates, a main effect emerged for Congruence, $F(1,46) = 13.13$, $p = .001$ (excluding neutral words from the analysis), but no other main effects or interactions approached significance. The mean number of errors per participant were low, as follows (out of a possible 20 per condition): M errors for congruent words = .44, $SD = .77$; M errors for incongruent words = .92, $SD = 1.33$; M errors for congruent names = .50, $SD = .77$; M errors for incongruent names = .85, $SD = 1.01$; M errors for neutral words = .73, $SD = .96$. When words were used as the unit of analysis, a main effect of Congruence again emerged, $F(1,38) = 5.23$, $p = .028$, but there was no other main effect or interaction.

Like the adults in Experiment 1, children in this study exhibited Stroop-like interference in response to incongruent trials. Interestingly, although adults had shown stronger interference to gender-stereotyped names than to gender-stereotyped words, children seemed to show the opposite pattern. A 2 (word vs. name) × 2 (adults vs. children) ANOVA performed on indices of the congruence effect (RTs for incongruent conditions minus RTs for congruent conditions; see Table 2) revealed that the difference in these patterns was significant, $F(1,84) = 8.77$, $p = .004$ (Two outliers were dropped from this analysis, but including them would only have strengthened this effect). This analysis only involved stimuli presented to both groups.[5] Further analyses on the effect of congruence revealed that adults were relatively more affected than children by congruence in response to gender-stereotyped names, $t(86) = 1.88$, $p = .064$, and children were more affected by congruence in response to gender-stereotyped words than adults were, $t(84) = 1.89$, $p = .063$. It is hard to say why adults should show more interference to incongruent names than to incongruent words, whereas children showed more interference to incongruent words than to incongruent names. One possibility is that the names used in this study were less familiar to the children. Such patterns may raise speculations about developmental changes (e.g., differences in gender flexibility surrounding words vs. names); however, the current data can only raise—not answer—such speculations.

General discussion

Across two experiments, adults and children who simply categorized the sex of a speaker's voice found it more difficult to do so when the spoken words were stereotypically incongruent with the speaker's gender. Similar to the classic Stroop task, where interference reflects an inability to disregard the meaning of stimuli, the current findings suggest that people who are engaged in a sex-based classification

[5] The adult RTs to only the congruent and incongruent stimuli also used with children were as follows: congruent words ($M = 776$ ms, $SD = 190$ ms), incongruent words ($M = 813$ ms, $SD = 200$ ms), congruent names ($M = 774$ ms, $SD = 190$ ms), and incongruent names ($M = 830$ ms, $SD = 223$ ms).

often can not disregard gender stereotyped associations even when they are task-irrelevant.

Of course, evidence for such associations does not necessarily suggest the strength or presence of personally held beliefs. Indeed, people are inundated daily with such volumes of gender-stereotypical information that it would not be surprising to find even the most explicitly egalitarian individual exhibiting Stroop-like interference in the present task. This leads to open empirical questions: might individual differences in socialization, culture, and development be reflected through interference on the auditory Stroop? Further, to what degree would interference in the task correlate with explicit measures such as those used in traditional indices of gender schema content? If the correlation is small, how might this measure differently predict overt behavior?

One interesting aspect of these two experiments is the fact that adults and children differed in the relative effects of congruence: for gender-stereotypical names, adults showed a somewhat greater effect of congruence than children did, and children showed a somewhat greater effect of congruence for words related to stereotyped gender roles. Although the current data cannot say anything about age-related changes, future investigations using this paradigm—testing a broader range of ages—might yield insights about the development of implicit gender schemas; children form gender schemas across many domains, including toys, activities, occupations, and traits, and gender schemas within these domains may follow different developmental trajectories (see Serbin et al., 1993). An auditory Stroop using words associated with such domains may usefully supplement the explicit measures commonly used to follow such trends (e.g., Edelbrock & Sugawara, 1978; Liben & Bigler, 2002; Liben, Bigler, & Krogh, 2002; Williams et al., 1975). Unlike the IAT—itself a powerful implicit tool (Greenwald et al., 1998)—the auditory Stroop does not require the explicit mapping together of categories and evaluative terms. The simplicity of the task suggests that it could be especially valuable for future research with children as young as three-years-old (see, for example, Jerger et al., 1988).

Although neither children nor adults in our samples appeared able to disregard task-irrelevant gender associations, one must be cautious about concluding that this reflects activation of gender stereotype upon merely hearing voices of different sexes. Indeed, the primary task involved a sex-categorization task, and this may have primed the activation of gender stereotypes, in which case the effect revealed here might not meet strict criteria defining "goal-independent" processing (see Bargh, 1989). Thus, it will be important to know whether similar interference would arise if the categorization task was independent of the sex of the speaker (e.g., is the voice high or low?). Based on the results of an earlier study using a different paradigm, we might expect such interference to be weaker but still present (Banaji & Hardin, 1996). For example, in that study, people were slower to respond to gender-related target pronouns (e.g., he, she) when they were preceded by stereotypically gender-incongruent primes (e.g., nurse, doctor) than by gender-congruent ones (Banaji & Hardin, 1996). This effect was strongest when the primary task was relevant to gender (i.e., is the pronoun masculine, feminine, or neutral) but persisted more weakly when the primary task was gender-irrelevant (i.e., is the target a pronoun or not), suggesting that implicitly measured gender-stereotype activation is enhanced by—but not dependent on—the nature of the task.

In conclusion, the Stroop-like effect elicited by voices saying words stereotypically incongruent with their own sex not only supports notions that the activation of irrelevant gender associations proceeds relatively automatically (at least when gender is a salient feature of a task), but it also yields a relatively direct task that is simple enough to be used with both adults and children. Investigations of the relationship between this effect and explicit measures of gender stereotyping, the degree to which it predicts overt behavior, and its modulation by factors such as gender role flexibility all promise to provide illuminating avenues for future research.

References

Banaji, M. R., & Hardin, C. (1996). Automatic gender stereotyping. *Psychological Science, 7*, 136–141.

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3–51). New York, NY: Guilford Press.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science, 17*, 53–58.

Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological Review, 88*, 354–364.

Biernat, M. (1991). Gender stereotypes and the relationship between masculinity and femininity: A developmental analysis. *Journal of Personality & Social Psychology, 61*, 351–365.

Bradbard, M. R., Martin, C. L., Endsley, R. C., & Halverson, C. F. (1986). Influence of sex stereotypes on children's exploration and memory: A competence versus performance distinction. *Developmental Psychology, 22*, 481–486.

Cedrus. (1996). *SuperLab Pro* (Version 1.05) [Computer Software]. Phoenix, AZ: Cedrus.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior, 12*(24), 335–359.

Cohen, G., & Martin, M. (1975). Hemisphere differences in an auditory Stroop test. *Perception & Psychophysics, 17*(1), 79–83.

Dovido, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology, 33*, 510–540.

Edelbrock, C., & Sugawara, A. I. (1978). Acquisition of sex-typed preferences in pre-school-aged children. *Developmental Psychology, 14*, 614–623.

Gauthier, I., Curran, T., Curby, K. M., & Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. *Nature Neuroscience, 6*(4), 428–432.

Glick, P., & Fiske, S. T. (1999). Gender, power dynamics, and social interaction. In M. M. Ferree, J. Lorber, & B. Hess (Eds.), *Revisioning gender: The gender lens* (Vol. 5, pp. 365–398). Thousand Oaks, CA: Sage Publications.

Green, E. J., & Barber, P. J. (1981). An auditory Stroop effect with judgements of speaker gender. *Perception & Psychophysics, 30*, 459–466.

Green, E. J., & Barber, P. J. (1983). Interference effects in an auditory Stroop task: Congruence and correspondence. *Acta Psychologica, 53*, 183–194.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Hamers, J. F., & Lambert, W. E. (1972). Bilingual interdependencies in auditory perception. *Journal of Verbal Learning and Behavior, 11*, 303–310.

Jerger, S., Martin, R. C., & Pirozzolo, F. J. (1988). A developmental study of the auditory Stroop effect. *Brain and Language, 35*, 86–104.

Katz, P. A., & Ksansnak, K. R. (1994). Developmental aspects of gender role flexibility and traditionality in middle childhood and adolescence. *Developmental Psychology, 30*(2), 272–282.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87*, 820–857.

Klein, G. S. (1964). Semantic power measured through the interference of words with color-naming. *American Journal of Psychology, 77*, 576–588.

Kuhn, D., Nash, S. C., & Brucken, L. (1978). Sex role concepts of two- and three-year-old children. *Child Development, 49*, 445–451.

Leinbach, M. D., & Fagot, B. I. (1993). Categorical habituation to male and female faces: Gender-schematic processing in infancy. *Infant Behavior and Development, 16*, 317–332.

Liben, L. S., & Bigler, R. S. (2002). The developmental course of gender differentiation: Conceptualizing, measuring, and evaluating constructs and pathways. *Monographs of the Society for Research in Child Development, 67*(2, Serial No. 269).

Liben, L. S., Bigler, R. S., & Krogh, H. R. (2002). Language at work: Children's gendered interpretations of occupational titles. *Child Development, 73*(3), 810–828.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*(4), 492–527.

Maccoby, E. E. (1988). Gender as a social category. *Developmental Psychology, 24*, 755–765.

Martin, C. L. (1999). A developmental perspective on gender effects and gender concepts. In W. B. Swann Jr., J. H. Langlois, & L. A. Gilbert (Eds.), *Sexism and stereotypes in modern society: The gender science of Janet Taylor Spence* (pp. 45–73). Washington, DC: American Psychological Association.

Martin, C. L., & Halverson, C. F., Jr. (1981). A schematic processing model of sex typing and stereotyping in children. *Child Development, 52*, 1119–1134.

Martin, C. L., & Ruble, D. (2004). Children's search for gender cues: Cognitive perspectives on gender development. *Current Directions in Psychological Science, 13*, 67–70.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the implicit association test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37*, 435–442.

Morgan, A. L. R., & Brandt, J. F. (1989). An auditory Stroop effect for pitch, loudness, and time. *Brain and Language, 36*, 592–603.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the implicit association test: II. Method variables and construct validity. *Personality & Social Psychology Bulletin, 31*, 166–180.

O'Brien, M., & Huston, A. C. (1985). Development of sex-typed play behavior in toddlers. *Developmental Psychology, 21*, 866–871.

Scheibe, K. E., Shaver, P. R., & Carrier, S. C. (1967). Color association values and response interference on variants of the Stroop test. *Acta Psychologica, 26*, 286–295.

Serbin, L. A., Powlishta, K. K., & Gulko, J. (1993). The development of sex typing in middle childhood. *Monographs of the Society for Research in Child Development, 58*(2, Serial No. 232).

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–661.

Syntrillium. (1996). *Cool Edit 96* [Computer Software]. Phoenix, AZ: Syntrillium.

Williams, J. E., Bennett, S. M., & Best, D. L. (1975). Awareness and expression of sex stereotypes in young children. *Developmental Psychology, 11*, 635–642.