Appendix A

fMRI Counterarguing Localizer Task

67 participants were recruited for the counterarguing localizer task as part of a larger study examining the effects of messages that promote physical activity (for more information about this sample, see Cooper, Tompson, O'Donnell, & Falk, 2015). To be eligible for the study, participants had to be right-handed, and meet standard fMRI criteria related to safety, including having no metal in their body (except for tooth filings), no history of psychiatric or neurological disorders, currently not taking any psychiatric or illicit drugs, were not claustrophobic, and were not pregnant or breast-feeding. The sample consisted of 41 females and 26 males, with a mean age of 33.42 years old (SD = 13.04). 44 of the participants were White, 12 Black, 3 Asian, 1 Hispanic and 7 others.

The stimuli used in the task were 70 generic behavioral statements starting with the same stem "People should..." (e.g. "People should do the crossword", "People should sing in the shower", "People should text while driving"), among which 31 were easy to agree with (i.e., more likely to argue in favor), 31 were hard to agree with (i.e., more likely to argue against), and 8 were in the middle. These statements were selected from a pool of statements which were generated and pretested on Amazon Mechanical Turk. The pre-test asked participants to generate as many as possible reasons "against" and reasons "in favor" to each of the statements in the pool. The difference score calculated by subtracting the numbers of reasons "against" and reasons "in favor" to a particular statement was used to determine whether the statement is easier to agree with or disagree with. A final set of 70 items, which consist of three types of statement – "easy in favor", "easy against", and "middle" – was selected from the larger pool based on the difference score for each statement. "Easy in favor" category includes statements: *adopt animals*

in need, always try to do better, continue to learn, do the crossword, follow the news, forgive others, help those in need, keep a journal, keep in touch with friends, learn another language, listen to others, listen to the radio, make new friends, make time for hobbies, reach goals, read more books, remember the past, respect elders, share with one another, sing in the shower, speak up, spend time with friends, take naps, take risks, take short showers, travel to other countries, use public transportation, volunteer, watch world news, work hard, work together. "Easy against" includes: act without thinking, arrive late, be late to appointments, bike on the sidewalk, bike without helmets, block an empty seat, boast about money, carve initials in trees, cheat on a test, draw on furniture, drive through red lights, drive too close together, get tattoos, go to work sick, hit other people, ignore current events, judge others, leave dishes in the sink, lie to friends, play loud music, put off deadlines, run with scissors, speed while driving, talk during movies, talk loudly on the phone, talk over another person, tell lies, text in meetings, text while driving, use plastic forks, write in library books. "Middle" includes get to bed early, sleep in, squash bugs and spiders, stay up late, talk to strangers, use weed killer, vote in local elections, wake up earlier.

The fMRI localizer task consisted of three within-subject conditions which were presented in a random order to the participants. In condition 1, participants were presented with 3 statements in each trial, and were instructed to make quick, gut level responses about whether they agreed or disagreed with the statements, and they only had 3 seconds to respond yes or no for each statement. In conditions 2 and 3, participants were presented one statement in each trial, and were asked to generate as many reasons in favor (condition 2) or against (condition 3) the statement over a period of 12 seconds as possible; they were also instructed to press a button with each reason they generated in the two conditions. Each condition had 10 trials, and had an

equal number of "easy in favor" and "easy against" statements randomly selected and ordered from the pool of 70 statements, to account for the level of difficulty in generating arguments in favor and against of the statements. Across conditions, each trial was followed by a fixation cross for 2 seconds.

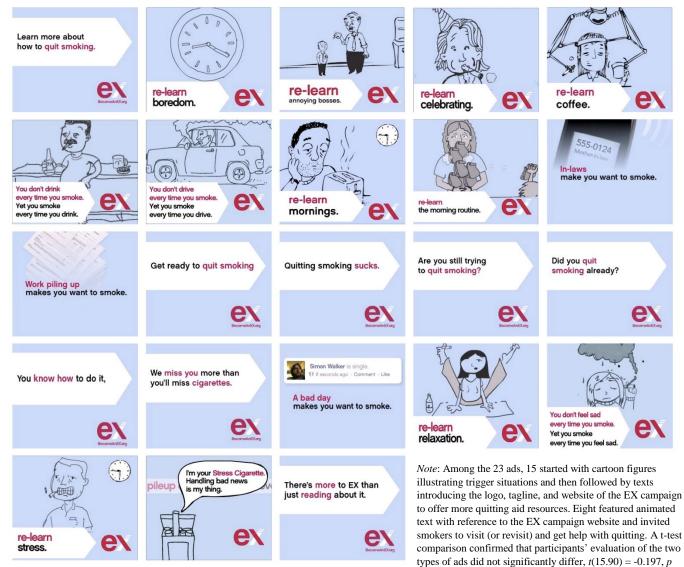
The neural regions of interests were localized through two sets of planned contrasts. To isolate neural systems that are associated with deliberative argumentation across participants, the first contrast examined conditions in which the participant was arguing both in favor and against (condition 2 & condition 3) versus condition 1 where the participant was asked to give quick response and not to deliberate. The resulting ROIs are the clusters in the anterior and bilateral dorsal lateral prefrontal cortex, or DLPFC ("deliberative argumentation ROI"; yellow areas highlighted in Figure 1 in the main text). The second comparison identified a cluster of voxels in the right posterior DLPFC that are most robustly associated with negative position in argumentation ("negative position ROI"; red areas highlighted in Figure 1 in the main text), specifically, as subjects were prompted to argue against (condition 3) versus in favor (condition 2) of several statements with deliberative processing. The findings of the localizer task are consistent with prior studies which have shown that neural activity in DLPFC is associated with effortful deliberation (Curtis & D'Esposito, 2003; Hutcherson, Plassmann., Gross, & Rangel, 2012; Rosenbloom, Schmahmann, & Price, 2012). These independently localized ROIs (O'Donnell, Coronel, Cascio, Lieberman, & Falk, 2018) were then used to investigate smokers' neural activity patterns during exposure to naturalistic anti-smoking messages in the main study.

UNDERLYING MECHANISMS OF COUNTERARGUING

Online Supplementary Materials

Appendix B

Ex Campaign Banner Ad Stimuli Used in the Study (Static Screenshots; N = 23)



= .85, reducing concerns about variability across ads.

Appendix C

fMRI Data Acquisition and Preprocessing

fMRI Data Acquisition. Neuroimaging data were acquired using a 3 Tesla GE Signa MRI scanner. One functional run of the banner ads task (304 volumes total) was acquired for each participant. Functional images were recorded using a reverse spiral sequence (TR = 2000 ms, TE = 30 ms, flip angle = 90°, 43 axial slices, FOV = 220 mm, slice thickness = 3mm; sequential descending slice acquisition; voxel size = $3.44 \times 3.44 \times 3.0$ mm). We also acquired inplane T1-weighted images (43 slices; slice thickness = 3 mm; voxel size = $.86 \times .86 \times 3.0$ mm) and high-resolution T1-weighted images (SPGR; 124 slices; slice thickness = $1.02 \times 1.02 \times 1.2$ mm) for use in coregistration and normalization.

fMRI Preprocessing. Functional data were pre-processed and analyzed using Statistical Parametric Mapping (SPM8, Wellcome Department of Cognitive Neurology, Institute of Neurology, London, UK). To allow for stabilization of the BOLD signal, the first five volumes (10s) of each run were discarded prior to collection. Functional images were despiked using the AFNI 3dDespike program (http://afni.nimh.nih.gov/afni). In SPM 8, data were corrected for differences in the time of slice acquisition using sinc interpolation, with the first slice serving as the reference. Spatial realignment of images was carried out with reference to the first functional image. We used two-stage co-registration to align functional and structural images: 1. In-plane T1 images were registered to the mean functional image; 2. The high-resolution T1 structural images were registered to the in-plane image. T1 images were then segmented and normalized to the skull-stripped MNI template provided by FSL ("MNI152_T1_1mm_brain.nii"). Lastly, functional images were smoothed using a Gaussian kernel (8 mm FWHM).

Appendix D

BOLD Signal and fMRI Data Analysis

fMRI assesses neuronal activity in brain regions by tracking differences in oxygenated and deoxygenated blood (i.e., the blood-oxygen-level-dependent [BOLD] signal), to determine whether the areas of interest are more or less active during a psychological task. This assessment relies on the principle that, compared to blood elsewhere, the blood flowing to an active region is more oxygenated, and thus has different magnetic properties. fMRI can detect active regions where blood is flowing to through mapping the spatial location of these magnetic properties (Lieberman, 2010, p.146). In this way, researchers can infer changes in neural activity during a task. This form of fMRI (i.e., BOLD imaging) has been used very commonly in communication and other social science research (for review, see Coronel & Falk, 2017; Lazar, 2008; Lieberman, 2010).

The raw fMRI data obtained were then subjected to standard preprocessing procedures, to make the data suitable for analysis. These procedures include despiking (to correct for noise and outliers), realignment (to correct for subjects' head movement), normalization (to put all subjects' brain scans into a single coordinate space so the brain structures can be compared across subjects who have inherently different underlying brain sizes and shapes), and spatial smoothing (averaging over adjacent voxels to increase the signal to noise ratio). See Lieberman (2010) p.146-147 for a thorough description of these standard pre-processing steps.

In an absolute sense, brain regions are constantly active (i.e., there is no absolute "stop", "rest" or "off" in a living brain). Typical fMRI studies gauge neural activity through comparisons, i.e., by contrasting differences between conditions during a scan (Coronel & Falk, 2017), but the units read out by an MRI scanner are arbitrary (i.e., do not follow an absolute scale across people and tasks, but rather are relative to the comparison in question). To standardize these units, relative

differences are often scaled by converting to "percent signal change" from one condition to another. In our study, the neural activity in the a priori determined ROIs was obtained by contrasting the BOLD signals in our focal ROIs when the participants were exposed to antismoking messages, versus the BOLD signals in these ROIs while participants were not being exposed to messages (i.e., during their rest/fixation periods). These continuous estimates of percent signal change from baseline to message exposure for each message are then compared relative to one another, when we correlated percent signal change in the ROIs for each message with continuous language scores based on the participants' descriptions of the messages after the scan.

Interested readers who would like to know more details about fMRI data analysis are referred to reviews by Coronel & Falk, 2017; Lazar, 2008; Lieberman, 2010; Sherry, 2015; Weber, 2015; Weber, Eden, et al., 2015; Weber, Fisher, Hopp, & Lonergan, 2018; and Weber, Mangus, & Huskey, 2015, for further reading.

Appendix E

Calculation of Simple Slopes

Following the procedure outlined in Cohen, Cohen, West, and Aiken (2003, p.564), we created two simple effect models: one for high intention to change and the other for low intention to change individuals. To do so, we first calculated the mean and SD of the moderator (i.e., the intention to change variable). We then estimated the simple effects of neural activity on language outcomes at high (M+1SD) vs. low (M-1SD) intention to change levels. We then created two simple effect models: one for high intention to change and the other for low intention to change individuals. The first simple effect model aims to estimate the effect of neural activity on language outcomes for individuals having higher intentions (M+1SD). To do that, we removed 1SD from cluster-mean centered intention to change. The second simple effect model aims to estimate the effect of neural activity on language outcomes for individuals having lower intentions (M-1SD). For this model, we added 1SD from cluster-mean centered intention to change. Although at first thought it might seem counter-intuitive to subtract in order to derive the effect at +1 SD, and vice versa for -1SD, the logic is that in each of the two simple effect models, we adjusted what zero meant such that the simple main effect was estimated when everything else was at zero. In the simple effect model for high "intenders", the coefficient estimate of neural activity on language outcomes is the simple slope of neural activity when intention to change = 0, i.e. in this case, when intention to change = +1 SD. And vice versa for the low "intenders".

Appendix F

Sensitivity Analyses

Alternative language outcomes. We conducted three sets of sensitivity analyses. The first set of sensitivity analyses focused on potential alternative language outcomes: 1) Breadth of *deliberative argumentation*: Another LIWC category "relativity", which consists of words that exhibit the level of recall specificity and describe details such as relative position, time and action, may to some degree reflect the breadth dimension as well. We thus examined this measure in our sensitivity analyses as a potential alternative proxy for the breadth dimension of individuals' deliberative argumentation; 2) Depth of deliberative argumentation: LIWC categories such as "words per sentence", or the "percent of words longer than six letters", may be indicative of more complex language use as well. However, these categories are often considered as general descriptor categories of individuals' linguistic style characteristics rather than manifestations of psychological constructs, and do not have clear established links with behavioral implications (Pennebaker et al., 2007). They may also reflect other individual differences such as tendency to use more filler words (e.g., blah, I mean, you know), or education levels, instead of revealing their message processing activities (Tausczik & Pennebaker, 2010). We examined these word categories in sensitivity analyses to determine whether these attributes were also relevant to our processes of interest, insofar as they might capture cognitive depth, with the caveat stated above; 3) Valence stance: The other focal language measure, negative position, was operationalized by extracting the valence stance or polarity of each text. In order to determine whether greater nuance in the valence dimension may better explain the phenomenon in question, we also examined the LIWC "positive emotion" category, as well as discrete negative emotion categories, including "anxiety", "anger", and

"sadness", in our sensitivity analyses.

The sensitivity analyses on alternative deliberation types that may emerge during message processing confirmed that the neural activity in the two hypothesized ROIs was not associated with non-argumentative or valence-neutral deliberation that emphasizes recall and description (i.e., <u>"relativity"</u>), or negative deliberation that features single discrete negative emotions such as "anger". This does not imply that these psychological processes are not at play during counterarguing, but rather that words associated with those processes in the LIWC dictionary are not specifically associated with brain activity in our primary counterarguing regions of interest. Further corroborating our findings related to the focal valence outcome (i.e., positivity score, which reflects valence dominance or polarity), we observed that greater neural activity within the "negative position" ROI was also indicative of less sheer amount of positive reflection in the descriptions (as quantified by the "positive emotion" category) among smokers who have lower intention to change.

Taken together, the sensitivity analyses results revealed that, these other alternative language outcomes were not associated with our hypothesized ROIs, with the exception of a significant negative conditional main effect observed for the "positive emotion" category, indicating that greater neural activity within the "negative position" ROI is indicative of fewer positive thoughts among smokers who have lower intention to change. See Table S2 (in online supplementary materials) for details of the results.

[Insert Table S2 here]

Gender and Education as Moderators. We also conducted sensitivity analyses by including gender and education as moderator variables in both main effects and interaction models, considering our sample contained more males (31 out of 44 smokers) and college

10

students (n = 21). Results suggested that gender and education were not significant moderators across all models (gender moderation effect: p-values range from .24 to .84; education moderation effect: p-values range from .10 to .90).

Multi-level Regressions Excluding Control Variables. To strengthen the findings and implications from the multi-level regression analyses which linked neural activity and language outcomes, we also performed sensitivity analyses by running the multi-level regressions again after excluding the control variables. The results are summarized in Table S4 of the online supplementary materials. As can be seen from Table S4, excluding the control variables did not affect the results of our study, and the patterns we observed still held stable.

[Insert Table S4 here]

Multi-level Regressions Excluding Message Evaluation Variable. We conducted a set of additional sensitivity analyses where the message evaluation variable was removed from all the models. The results are summarized in Table S5 of the online supplementary materials. As can be seen from Table S5, the main results and conclusions were similar with or without controlling for self-reports. We thus confirmed that neural activity during initial exposure to stimuli could predict additional variance in participants' subsequent reactions towards the messages above and beyond self-report measures of the message effectiveness evaluation.

[Insert Table S5 here]

In sum, these additional sets of sensitivity analyses further corroborated the robustness of the study findings.

Appendix G

Intention to Change Measure

As shown in Figure S1, although we screened for smokers who were not immediately planning to quit, we observed significant variability in their intention to change their behavior in the next three months/ openness to changing some aspects of their behavior (even if not fully quitting). Through this measure, we can further distinguish those who were more determined to not take any actions about their smoking behavior in the foreseeable future, with those who were less adamant and may have already recognized that their smoking behavior can produce negative consequences. We expected that the extent of counterarguing would be stronger for the smokers who had the lowest intention to change and were most committed to their smoking behavior, and hence most likely to defend their smoking behavior and negatively react to the anti-smoking messages.

[Insert Figure S1 here]

The mean score of the intention to change measures was 2.42 (range = 1 - 4; *SD* = 0.81), suggesting on average low to moderate intention to reduce or abstain from smoking in the next three months. As can be seen from the histogram of the intention to change variable in Figure S1, although more participants were on the lower end of the intention to change composite score, there were still several participants who had relatively moderate or high scores on this measure. We consider this variable to have sufficient variation in our sample.

Table S1

	by Person $(N = 44)$							by A	d (N	7 = 23)	
	1	2	3	4	5	6	1	2	3	4	5	6
1-Deliberative argumentation ROIs												
2-Negative position ROI	.66***						.60**					
3–Intention to change	14	18										
4–Word count	31*	10	.02				44*	42*				
5–Cognitive mechanism	29	13	.10	04			.16	.29		26		
6–Positivity	22	06	13	20			.08	.09		10	25	

Zero-order Correlations of Focal Variables at Person and Ad Levels

Note. Pairwise Pearson's correlation coefficients are presented. * p < .05, ** p < .01, *** p < .001.

UNDERLYING MECHANISMS OF COUNTERARGUING

Online Supplementary Materials

Table S2.

Neural Activity in the Functionally Localized Deliberative Argumentation and Negative Position ROIs Interacting with Intention to Change on Alternative Language Measures, Controlling for Evaluation of Ads, Nicotine Dependence Level and Individuals' Average Word Count

DVs	Word p	per Sentence	Words	>6 letters	Rela	ativity	Positive	e Emotion	An	xiety	А	nger	Sa	dness
IVs	Main	Int.	Main	Int.	Main	Int.	Main	Int.	Main	Int.	Main	Int.	Main	Int.
Neural activity in deliberative argumentation ROIs	018	017	.009	.006	.048	.048								
Neural activity in negative position ROI							038	055	.024	.024	022	021	007	007
Intention to change	104	108	059	051	.022	.024	137	106	.021	.022	029	030	023	021
Evaluation of ads	.014	.013	062	061	.096*	.096*	.105**	.108**	011	011	059	059	.007	.007
Nicotine dependence	164	165	190*	190*	.075	.075	145	139	033	032	.127	.127	.011	.011
Average word count	.270*	.270*	057	056	.183**	.183**	223*	216*	008	008	150	150	.061	.062
Deliberative argumentation ROIs \times Intention to change		019		.042		.013								
Negative position ROI \times Intention to change								.121***		.004		003		.006

Note. Int. = Interaction Model. % > 6 letters refers to percent of words longer than six letters. Standardized multilevel regression coefficients γ are reported in this table. Main effects models do not include an interaction term. Interaction models contain the interaction between neural activity and intention to change. Variables involved in interactions were all mean-centered before entering the regression models. * p < .05, ** p < .01, *** p < .001.

UNDERLYING MECHANISMS OF COUNTERARGUING

Online Supplementary Materials

Table S3.

Models	Deliberative argumentation ROIs	Negative position ROI	Intention to change	Evaluation of ads	FTND	Verbosity	Interaction
DV=	1.012		1.095	1.048	1.002	1.008	1.050
Word Count							
DV = Cognitive Mechanism	1.008		1.052	1.030	1.002	1.005	1.027
DV = Positivity		1.021	1.019	1.010	1.002	1.001	1.031

Variance Inflation Factors (VIF) of the Variables included in the Multi-Level Regression Models

Note. Multicollinearity among predictor variables was assessed using VIF (variance inflation factor). The results suggested low VIF values across all variables in all models (ranging from 1.00-1.10), indicating multicollinearity was not a concern in all models.

Table S4.Multi-Level Regression Analyses Results After Excluding the Control Variables

DVs	Word	Word Count		Mechanism	Positivity		
IVs	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	
Neural activity in deliberative argumentation ROIs	-0.030	-0.031	0.036	0.039			
Neural activity in negative position ROI					-0.029	-0.037	
Intention to change	0.015	0.019	0.029	0.011	-0.072	-0.053	
Deliberative argumentation ROIs \times Intention to change		0.016		-0.080*			
Negative position ROI $ imes$ Intention to change						0.087**	

Note. Standardized multilevel regression coefficients γ are reported in this table. Models 1, 3 and 5 present main effects models without the interaction term. Models 2, 4 and 6 present conditional main effects models where interaction between neural activity and intention to change are taken into consideration. Variables involved in interactions were all mean-centered before entering the regression models. * p < .05, ** p < .01, *** p < .001.

Table S5.

	-			
Conditivity Analysis	Without	Controlling	for Evaluation	a of Ada
Sensitivity Analysis	winoui	Controlling	τοι εναιματιο	n or Aas
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~			<i>J</i> = <i>i</i> =	

DVs	Word	Count	Cognitive l	Mechanism	Positivity		
IVs	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	
Neural activity in deliberative argumentation ROIs	028	.001	0.04	0.046			
Neural activity in negative position ROI					-0.022	-0.034	
Intention to change	.016	.012	-0.002	-0.055	-0.067	-0.013	
Nicotine dependence	004	007	-0.052	-0.058	-0.153	-0.143	
Average word count	.789***	.790***	0.006	0.001	-0.104	-0.095	
Deliberative argumentation		020		0.070*			
$ROIs \times Intention$ to change		030		-0.079*			
Negative position ROI $\times$						0.092**	
Intention to change						0.092	

*Note*. Standardized multilevel regression coefficients  $\gamma$  are reported in this table. Models 1, 3 and 5 present main effects models without the interaction term. Models 2, 4 and 6 present conditional main effects models where interaction between neural activity and intention to change are taken into consideration. Variables involved in interactions were all mean-centered before entering the regression models. * p < .05, ** p < .01, *** p < .001.

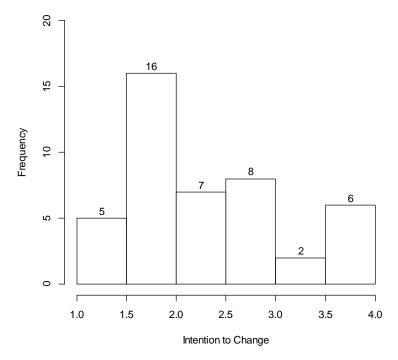


Figure S1. Frequency Distribution of Intention to Change

#### References

- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd edition). Mahwah, N.J: Routledge.
- Cooper, N., Tompson, S., O'Donnell, M. B., & Falk, E. B. (2015) Brain activity in self-and value-related regions in response to online antismoking messages predicts behavior change. *Journal of Media Psychology*, 27, 93-108. doi: 10.1027/1864-1105/a000146
- Coronel, J. C., & Falk, E. B. (2017). Functional magnetic resonance imaging and communication science. In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *The international encyclopedia* of communication research methods. Hoboken, NJ: John Wiley & Sons, Inc.
- Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, 7, 415–423. doi: 10.1016/s1364-6613(03)00197-9
- Hutcherson, C. A., Plassmann, H., Gross, J. J., & Rangel, A. (2012). Cognitive regulation during decision making shifts behavioral control between ventromedial and dorsolateral prefrontal value systems. *Journal of Neuroscience*, *32*, 13543–13554. doi: 10.1523/JNEUROSCI.6387-11.2012
- Lazar, N. (2008). *The statistical analysis of functional MRI data*. New York, NY: Springer-Verlag.
- Lieberman, M. D. (2010). Social cognitive neuroscience. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology*. Hoboken, NJ: John Wiley & Sons, Inc.
- O'Donnell, M. B., Coronel, J., Cascio, C. N., Lieberman, M. D., & Falk, E. B (2018, May). An *fMRI localizer for deliberative counterarguing*. Paper presented at The Social & Affective Neuroscience Society Annual Meeting, Brooklyn, NY.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC.Net.

Rosenbloom, M. H., Schmahmann, J. D., & Price, B. H. (2012). The functional neuroanatomy of decision-making. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 24(3), 266–277. doi: 10.1176/appi.neuropsych.11060139

- Sherry, J. L. (2015). Neuroscience and communication. *Communication Methods and Measures*, 9, 117–122. doi: 10.1080/19312458.2014.999756
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 24–54. doi: 10.1177/0261927X09351676
- Weber, R. (2015). Biology and brains methodological innovations in communication science: Introduction to the special issue. *Communication Methods and Measures*, 9, 1–4. doi: 10.1080/19312458.2014.999755
- Weber, R., Eden, A., Huskey, R., Mangus, J. M., & Falk, E. (2015). Bridging media psychology and cognitive neuroscience. *Journal of Media Psychology*, 27, 146–156. doi: 10.1027/1864-1105/a000163
- Weber, R., Fisher, J. T., Hopp, F. R., & Lonergan, C. (2018). Taking messages into the magnet: Method–theory synergy in communication neuroscience. *Communication Monographs*, 85, 81–102. doi: 10.1080/03637751.2017.1395059
- Weber, R., Mangus, J. M., & Huskey, R. (2015). Brain imaging in communication research: A practical guide to understanding and evaluating fmri studies. *Communication Methods* and Measures, 9, 5–29. doi: 10.1080/19312458.2014.999754