Automatic detection of Autism Spectrum Disorder in children using acoustic and text features from brief natural conversations

Sunghye Cho[1], *Mark Liberman*[1], *Neville Ryant*[1], *Meredith Cola*[2], *Robert T. Schultz*[2,3], *and Julia Parish-Morris*[2,4]

[1]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA
[2]Center for Autism Research, Children's Hospital of Philadelphia, Philadelphia, PA
[3]Department of Pediatrics, University of Pennsylvania, Philadelphia, PA
[4]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA

csunghye@sas.upenn.edu, myl@cis.upenn.edu, nryant@ldc.upenn.edu, colam@email.chop.edu, schultzrt@email.chop.edu, parishmorrisj@email.chop.edu

Abstract

Autism Spectrum Disorder (ASD) is increasingly prevalent [1], but long waitlists hinder children's access to expedient diagnosis and treatment. To begin addressing this problem, we developed an automated system to detect ASD using acoustic and text features drawn from short, unstructured conversations with naïve conversation partners (confederates). Seventy children (35 with ASD and 35 typically developing (TD)) discussed a range of generic topics (e.g., pets, family, hobbies, and sports) with confederates for approximately 5 minutes. A total of 624 features (352 acoustic + 272 text) were incorporated into a Gradient Boosting Model. To reduce dimensionality and avoid overfitting, we dropped insignificant features and applied feature reduction using Principal Component Analysis. Our final model was accurate substantially above chance levels. Predictive features were both acoustic-phonetic and lexical, from both participants and confederates. The goal of this project is to develop an automatic detection system for ASD that relies on very brief, generic, and natural conversations, which can eventually be used for ASD prescreening and triage in real-world settings such as doctor's offices and schools.

Index Terms: clinical speech, machine learning, Autism Spectrum Disorder

1. Introduction

The earliest descriptions of autism spectrum disorder (ASD) include mention of atypical speech patterns [2], [3]. Linguistic production has been explored in ASD, but most prior research samples were either elicited in highly structured contexts (e.g., reading sentences or word lists; [4]) or drawn from semi-structured clinical interviews with an autism expert (i.e., Autism Diagnostic Observation Schedule (ADOS) evaluations [5]; [6]). While valuable, these studies produce results that may not generalize to the everyday conversations that impact the lives of children on the autism spectrum. In this study, we address a gap in the literature by developing and testing machine learning classification approaches for characterizing children's short natural interactions with a non-expert conversational partner.

Numerous speech and language features distinguish children with ASD from TD children, including word choice and prosody. In recent years, researchers have begun to use those features for automatic classification using machine-learning techniques. For example, [7] measured pitch features from ADOS interviews of 146 children, including mean and median F0 values and median absolute deviation from the median (MAD), and trained a Naïve Bayes classifier using leave-one-out cross validation. Results showed that this approach correctly classified samples from ASD and TD children approximately 74% of the time, suggesting that pitch features are useful for identifying ASD. Similarly, [8] examined "awkward" prosody in 43 children with ASD and 26 TD controls using semi-structured data drawn from a story retelling task. They measured speech rate and rhythm, voice quality, and other intonational features. Results revealed that the model trained on speech rate and rhythm features performed the best, correctly classifying children with ASD and TD approximately 69% of the time. These results, while promising, were drawn from controlled samples that may not generalize to the real world.

Prior research also suggests that speech and language features produced by conversational partners (e.g., psychologists during clinical assessments) can be used to predict an autism diagnosis. For example, [9] examined speech features produced by 28 children and psychologists during a semi-structured clinical interview (ADOS). Results revealed that differences in both conversational partners' voices increased as ASD symptoms became more severe. Interestingly, psychologists' speech features predicted children's autism symptom severity better than child-based features. Concurrent evidence for the importance of interlocutor features comes from [10], who found that psychologists' speech features better predicted children's level of engagement in semi-structured dyadic interactions than children's features.

Taken together, the extant literature suggests that (1) child speech features can be used to classify ASD and (2) adding features from conversation partners can improve classification accuracy. However, while valuable, most previous studies used elicited speech samples or semi-structured clinical conversations, which are costly to collect and may not generalize to daily life. In this study, we address a gap in the literature by developing and testing machine learning classification approaches to children's natural interactions with a naïve conversational partner.

2. Objectives

Our objectives are to (i) automatically extract speech and text features that are predictive of ASD or TD status, (ii) train and

evaluate a predictive model, experiment with the extracted features, and (iii) identify the most predictive speech and text features for classifying ASD.

3. Methods

3.1. Participants

Seventy children with ASD (N=35, 13 females) or TD (N=35, 11 females) completed a 5-minute "get-to-know-you" conversation with a novel young adult confederate (N=22, 19 females). Diagnoses were confirmed (ASD group) or ruled out (TD group) using the Clinical Best Estimate process [11] informed by the Autism Diagnostic Observation Schedule – Second Edition (ADOS-2; [5]) and adhering to DSM-V criteria for ASD [12]. Groups were matched on Full Scale IQ estimates (Wechsler Abbreviated Scale of Intelligence – 2nd Edition; [13]), verbal and nonverbal IQ estimates, and sex ratio (Table 1). Participant social and repetitive behavior symptoms were characterized using ADOS-2 Calibrated Severity Scores [14] and scores on the Social Communication Questionnaire (SCQ; [15]). All participants were native English speakers.

Table 1: *Demographic and clinical characteristics of the participants. Shaded rows indicate clinical measurements.*

	ASD	**TDC**	**Group difference**
No.	35	35	
Age (sd)	11.42 (2.51)	10.57 (2.82)	$t = 1.33$, $p = 0.19$
Sex	13 f., 12 m.	11 f., 14 m.	$\chi^2 = 0.06$, $p = 0.8$
IQ	105.51	107.14	$t = -0.53$, $p = 0.6$
Verbal	104.97	105.83	$t = -0.28$, $p = 0.78$
Non-verbal	104.63	105.97	$t = -0.42$, $p = 0.68$
ADOS-2 overall	6.57	1.23	$t = 13.78$, $p < 0.001$
Soc. aff.	6.83	1.71	$t = 13.15$, $p < 0.001$
RRB	6.54	1.57	$t = 10.53$, $p < 0.001$
SCQ lifetime	17.68	2.69	$t = 10.99$, $p < 0.001$

3.2. Procedure

This study was conducted at the Center for Autism Research (CAR) with approval from and oversight by the Institutional Review Board of the Children's Hospital of Philadelphia (CHOP). Parents provided written informed consent for their minor children to participate, and children provided verbal assent. Language samples were drawn from a 5-minute unstructured conversation between children and a young adult confederate; conversational prompts were not provided to either speaker, they were simply instructed to "get to know each other". Confederates were unaware of participants' diagnostic status and were assigned to each participant based on availability. Conversations were audio/video recorded using a device (Biosensor) placed on a table between speakers. The Biosensor is equipped with two HD video cameras facing opposite directions, so that recordings of the participant and confederate were simultaneously captured as they sat facing each other during the conversation. Audio was recorded via four directional microphones embedded in the Biosensor.

3.3. Processing and Annotation

A team of reliable annotators produced time-aligned, verbatim, orthographic transcripts of audio recordings in XTrans [16]. Each recording was processed by two junior annotators and one senior annotator, all of whom were undergraduate students and native English speakers. Before becoming junior annotators for this cohort, each team member received at least 10 hours of training in Quick Transcription [17] modified for use with clinical interviews of participants with ASD. In addition, annotators achieved reliability (defined as > 90% in common with a Gold Standard transcript) on segmenting (marking speech start and stop times) and transcribing (writing down words and sounds produced, using the modified Quick Transcription specification) before beginning independent annotation. One reliable junior annotator segmented utterances into pause groups, while the second transcribed words produced by each speaker. A senior annotator with at least 6 months of annotation experience then thoroughly reviewed and corrected each file.

3.4. Features

3.4.1. Audio features

We converted the audio recordings from .flac to .wav using *sox*. Audio feature extraction was modeled after [18] using the ComParE13 configuration file of openSMILE [19] with a 25 ms window size and a 10 ms step size. Forty-four low-level descriptors (LLD) were measured by the configuration file, including voicing probability, pitch, jitter (local), jitter (DDP), shimmer (local), Harmonic-to-Noise Ratio (HNR), Root-Mean-Square energy (RMS), Mel Frequency Cepstral Coefficients (MFCC) from 1st to 14th, and zero crossing rate (ZCR). For each extracted feature, we also considered first-order differences, which are indicated by the suffix Δ (e.g., RMS-Δ). Frames containing overlapping speech and speech from the research assistants proctoring the conversation were removed, as were voiceless frames. Voiceless frames were defined as frames where the pitch value was 0 or voicing probability was more than 0.1 standard deviation below the mean for each speaker in each conversation.

We first z-normalized energy-related features, and after excluding voiceless frames, we normalized pitch values in Hz to semitones, using the 5th percentile of each speaker's pitch range as a baseline ($St = \log2(f0/baseline)*12$). Missing values were imputed using the SimpleImputer function in *scikit-learn* [20] in Python.

Since we were interested in speaker-level features for ASD classification, for each conversation we processed the participant and confederate LLD features using four statistical functionals: mean, median, standard deviation, and interquartile range (IQR = 75th percentile – 25th percentile). The total number of acoustic features generated by openSMILE was 352 (= 44 LLDs * 2 speakers * 4 functionals).

3.4.2. Text features

Text features were calculated using base R, qdap [21], Linguistic Inquiry and Word Count (LIWC) software [22], and a script written by one of the authors (ML) to measure word-related features, such as the frequencies of total words per speakers, pronoun usages, the number of filler words (um and uh). There were six main feature groups: pause/overlap metrics, segment/turn metrics, speaking rate/word complexity metrics, LIWC categories, lexical entropy/diversity measures, and parts

of speech. Formality and polarity were also computed at the conversation level for each speaker, using all words produced by a given speaker, leading to a total of 272 linguistic features (136 x 2 speakers).

3.5. Feature selection

Given the limited number of children in our sample, training our classifier with 624 features (352 acoustic features + 272 text features) is likely to overfit, resulting in poor performance. To reduce dimensionality and promote effective learning, we first selected features significantly correlated with the diagnostic status of the training set ($p < 0.025$) using univariate Pearson correlation within each cross-validation fold. This process selected 24–39 features per fold, for a total of 60 features (Table 3). 24–39 features for 70 samples are still likely to produce an overfitted model, so we further implemented Principal Component Analysis for feature reduction and used the first 10 components for model training.

Table 2: *Features that were significantly correlated with diagnostic status in the training set and selected more than 50% of the time. The number in parentheses indicates how many times the feature was selected (out of 70 folds).*

Participant (17 features)	
Acoustic features:	IQR & SD of voicing probability (70), Median of HNR-Δ (70), IQR of RMS-Δ (70), Median of 6th and 7th MFCC-Δ (70), IQR of HNR (67), Mean of 3rd MFCC-Δ (61), Mean and Median of 6th MFCC (55, 46), Median ZCR (44)
Text features:	Overlap percent (70), See-related words (70), Perception-related words (70), Relative words (66), Death-related words (61)
Confederate (16 features)	
Acoustic features:	Median, mean, IQR, SD of jitter DDP (70), Mean, SD, IQR of jitter local (70), Mean of jitter local-Δ (70), IQR of F0-Δ (70), Median of HNR-Δ (70), SD and Mean of 1st MFCC-Δ (55), Median of jitter local (50), Mean and SD of F0 (40)

Other selected features not shown in Table 3 include lexical features, such as words indicating informality, achievement, and rewards, as well as acoustic features, such as the median and IQR of both speakers' F0-Δ. Selected features were trained in a Gradient Boosting Classifier using *scikit-learn* in Python. We implemented leave-one-out cross validation to evaluate the generalizability of our model.

4. Feature analysis

Figure 1 illustrates some of the voice quality features that were selected in all CV folds (Table 3). For example, a linear regression model shows that the median HNR-Δ of ASD children is higher than TDC ($t = 3.25$, $p = 0.002$) and the same feature of confederates is higher when talking to ASD than to TDC ($t = 2.72$, $p = 0.008$; Fig. 1).

We also found several other features of ASD children that showed significantly lower values than TDC. For example, the standard deviation of participants' voicing probability ($t = 2.92$, $p = 0.004$) and IQR of F0-Δ ($t = 3.09$, $p = 0.003$) were significantly lower in ASD than TDC. Also, we found that some voice features of confederates were lower when talking to ASD children than TDC, such as all functionals of jitter (e.g., median jitter: $t = 2.7$, $p = 0.009$).

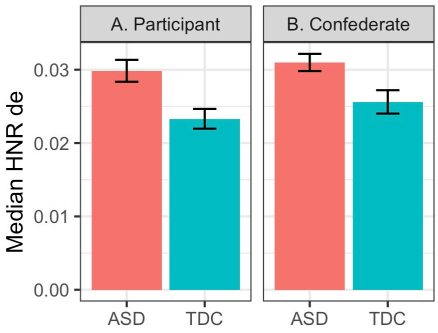

Figure 1: *The median Harmonic-to-Noise Ratio-Δ feature by speaker.*

5. Human perception

Before evaluating our model, we tested whether humans could identify speaker diagnosis based on brief natural conversations. We recruited two undergraduate students from the University of Pennsylvania to classify a subset of 12 conversations, where six of the participants were children with ASD and six were TD participants individually matched on age, sex, and IQ. The students were not experts in linguistics or speech/language pathology, but had worked at the Center for Autism Research for approximately 6 months. They were asked to listen to conversations one time through (without watching the video), and then guess the child's diagnostic category. In addition, we asked students to note the features that influenced their decision (per each conversation; Table 4). The comments in Table 4 indicate that a human's perception of ASD is related to atypical intonation, long and frequent pauses, relatively brief responses, and also pragmatic aspects of the conversations, such as topicality, relevance, and appropriateness of content.

Student A correctly identified the correct diagnostic category for 11 out of 12 participants (accuracy: 91.7%), whereas Student B had 9 correct answers (accuracy: 75%). Student A incorrectly identified one ASD as TD (false negative), and Student B misidentified two TD as ASD (false positive) and one ASD as TD (false negative). Students both classified the same ASD participant as a TD child, suggesting that our data realistically represent the ASD population, wherein participant phenotype can be highly heterogeneous. The students' mean accuracy was 83.33% (= 20/24 * 100).

Table 4: Factors influencing student decisions about diagnostic classification (only from correct answers).

ASD	• Atypical (flat-sounding) intonation • Frequent use of "I don't know" • Long latency to respond • Brief responses • Irrelevant details • Many random pauses • No follow-up questions • Less reactions • Changes in topic • Not much verbalization

TDC	• Relevant subject matter • Appropriate reactions • Typical intonation • Elaborate, timely responses • Many questions to the confederate • No abrupt changes in topic • Natural speech pause lengths • Few pauses • Good conversational flow

6. Classification results

As shown in Table 5, our classifier correctly identified the diagnostic status of the children 75.71% of the time. For TD children, the classifier was correct 85.71% of the time (30 out of 35). For ASD children, accuracy fell to 65.71% (23 out of 35). The classifier's receiver operator characteristic (ROC) curve is displayed in Figure 2, which is, again, far above the chance level (AUC = 75.43%).

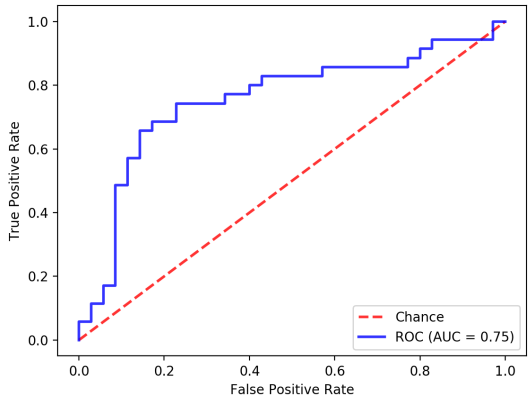

Figure 2: *Receiver Operator Characteristic Curve.*

Table 5: *Classification report of the model (top) and human performance (bottom). Note that human and system performance are not measured identically.*

		Accuracy	Precision	Recall	F1-score
System:	ASD	0.66	0.82	0.66	0.73
	TD	0.86	0.71	0.86	0.78
	Mean	0.76	0.76	0.76	0.76
Human:	ASD	0.67	0.73	0.89	0.80
	TD	1.00	1.00	0.80	0.89
	Mean	0.83	0.86	0.84	0.84

Importantly, our model was similarly accurate for classifying ASD (66%) compared with human raters (67%). However, human raters outperformed our system in classifying TD children. The model's F1-score was higher for TD than for ASD, meaning that it predicted TD children better than ASD overall. This seems to be because ASD patients have variable and heterogeneous symptoms rather than common, shared symptoms which makes predicting ASD harder than predicting TD children. The same trend was also found in human raters' performance. Lastly, our model had a high precision for classifying ASD, suggesting that it predicted ASD in a relatively conservative way.

We also experimented with different sets of features to investigate the most predictive features, including demographic information such as age and gender of the participants. However, adding age and gender to the model did not improve the performance and adding IQ-related measures rather worsened the performance. This seems to be because participants in our data were purposefully matched on age, gender, and IQ.

7. Discussion and Conclusion

This paper reports the results of an automatic classification system for ASD using features drawn from short natural conversations, where 35 ASD and 35 TD children discussed a range of general topics with a novel conversation partner. We extracted 624 acoustic and text features from children and their conversation partners, which were then reduced to 10 dimensions by performing feature selection and retaining the top 10 principal components within each CV fold. Our final model correctly classified ASD and TD children 75.71% of the time. This performance is reasonable, given that conversational samples were only 5 minutes long and topics were generic compared to semi-structured clinical interviews designed to elicit ASD-like features (e.g., ADOS evaluations). Student ratings of a subset of samples revealed that classifying ASD using these short, natural conversations is not an easy task, showing only slightly higher accuracy (83.33%) than our current model. However, we believe that using short, natural conversations like ours to classify ASD is a valuable endeavor, given that data collection is relatively easy and cheap and results are generalizable; therefore, it could potentially be applied to prescreen ASD in community-based settings. However, it should be noted that this approach is not yet valuable as a method for diagnosing ASD, which still needs to be diagnosed by expert clinicians.

Our classifier performed above chance, but the fact that student ratings were more accurate than our model suggests that humans attend to features that are not yet captured by our automated approach. For example, students noted topicality, relevance, appropriateness, and conversational flow as important factors influencing their diagnostic decisions. Our feature set indirectly measured these features (e.g., via LIWC categories) but did not comprehensively capture these dimensions on a turn-by-turn level. In future research, we plan to engineer features that more directly measure the semantic and pragmatic appropriateness of natural conversation samples. Also, we aim to increase the size of our dataset, with the goal of improving model performance and eventually applying this approach in real-world settings. Lastly, we note that longer conversations (e.g., 10 minutes instead of 5 minutes), and designated topics might improve model performance.

8. Acknowledgements

We thank the children and families that participated in our research, as well as clinicians and staff at the Center for Autism Research. This study was supported by an Autism Science Foundation Postdoctoral Fellowship to JPM; the Eagles Charitable Trust, McMorris Family Foundation, and Allerton Foundation to RTS; and NICHD 5U54HD086984-03 to Michael B. Robinson & RTS.

9. References

[1] J. Baio, "Prevalence of Autism Spectrum Disorder among children aged 8 years – Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014," *MMWR. Surveillance Summaries,* vol. 67, no. 6, pp. 1–23, 2018.

[2] H. Asperger, "Die Autistisehen Psychopathen im Kindesalter," *Arch. Psych. Nervenkrankh,* vol. 117, pp. 76–136, 1944.

[3] L. Kanner, "Autistic disturbances of affective contact," *Nerv. Child,* vol. 2, pp. 217–250, 1943.

[4] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders," in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25–29, Lyon, France, Proceedings,* 2013, pp. 191–194.

[5] C. Lord, M. Rutter, P. S. DiLavore, S. Risi, K. Gotham, and S. L. Bishop, *Autism diagnostic observation schedule, second edition (ADOS-2).* Torrance, CA: Western Psychological Services.

[6] J. Parish-Morris, M. Liberman, N. Ryant, C. Cieri, L. Bateman, E. Ferguson, and R. T. Schultz, "Exploring autism spectrum disorders using HLT," in *Proceedings of North American Association of Computational Linguistics, Comp Ling and Clin Psych,* pp. 74–84, 2016.

[7] G. Kiss, J. Van Santen, E. Prud'Hommeaux, and L. M. Black, "Quantitative analysis of pitch in speech of children with neurodevelopmental disorders," in *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association, September 9–13, Portland, OR, USA, Proceedings*, 2012, pp. 1342–1345.

[8] D. Bone, M. P. Black, A. Ramakrishna, R. Grossman, and S. Narayanan, "Acoustic-prosodic correlates of 'awkward' prosody in story retellings from adolescents with autism," in in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany, Proceedings*, 2015, pp. 1616–1620.

[9] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "The Psychologist as an interlocutor in Autism Spectrum Disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearning Research,* vol. 57, pp. 1162–1177, 2014.

[10] R. Gupta, D. Bone, S. Lee, and S. Narayanan, "Analysis of engagement behavior in children during dyadic interactions using prosodic cues," *Computer Speech and Language,* vol. 37, pp. 47–66, 2016.

[11] C. Lord, E. Petkova, V. Hus, W. Gan, F. Lu, D. M. Martin, O. Ousley, L. Guy, R. Bernier, J. Gerdts, et al., "A multisite study of the clinical diagnosis of different autism spectrum disorders," *Archives of General Psychiatry,* vol. 69, no. 3, pp. 306–313, 2012.

[12] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders, 5th edition: DSM-5.* Washington, D.C.: American Psychiatric Publishing, 2013.

[13] D. Wechsler, "Wechsler Abbreviated Scale of Intelligence–Second Edition (WASI-II). San Antonio, TX: NCS Pearson.

[14] K. Gotham, A. Pickles, and C. Lord, "Standardizing ADOS scores for a measure of severity in autism spectrum disorders," *Journal of Autism and Developmental Disorders,* vol. 39, no. 5, pp. 693–705, 2009.

[15] M. Rutter, A. Bailey, and C. Lord, *SCQ: The Social Communication Questionnaire*, Los Angeles: Western Psychological Services, 2003.

[16] Linguistic Data Consortium, "XTrans: A next generation translation tool developed by LDC [Online]," Philadelphia, PA, USA, 2009, Available: https://www.ldc.upenn.edu/language-resources/tools/xtrans

[17] M. L. Glenn, S. M. Strassel, H. Lee, K. Maeda, R. Zakhary, and X. Li, "Transcription methods for consistency, volume, and efficiency," in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, May 17–23*, Valletta, Malta, 2010, pp. 2915–2920.

[18] T. Alhanai, R. Au, and J. Glass, "Spoken language biomarkers for detecting cognitive impairment," *IEEE Automatic Speech Recognition and Understanding Workshop,* pp. 409–416, 2017.

[19] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proceedings of ACM Multimedia, Barcelona, Spain, ACM,* 2013, pp. 835–838.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825–2830, 2011.

[21] T. W. Rinker, "qdap: Quantitative Discourse Analysis Package 2.3.2," Buffalo, NY, 2019.

[22] Y. R. Tausczik & J. W. Pennebaker. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676