Accepted Manuscript

Eavesdropping on Character: Assessing Everyday Moral Behaviors

Kathryn L. Bollich, John M. Doris, Simine Vazire, Charles L. Raison, Joshua J. Jackson, Matthias R. Mehl

PII: S0092-6566(15)30031-3
DOI: http://dx.doi.org/10.1016/j.jrp.2015.12.003
Reference: YJRPE 3512

To appear in: *Journal of Research in Personality*

Received Date: 8 December 2015
Accepted Date: 23 December 2015

Please cite this article as: Bollich, K.L., Doris, J.M., Vazire, S., Raison, C.L., Jackson, J.J., Mehl, M.R., Eavesdropping on Character: Assessing Everyday Moral Behaviors, *Journal of Research in Personality* (2015), doi: http://dx.doi.org/10.1016/j.jrp.2015.12.003

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Eavesdropping on Character: Assessing Everyday Moral Behaviors

Kathryn L. Bollich[1*], John M. Doris[2], Simine Vazire[3], Charles L. Raison[4], Joshua J. Jackson[1], & Matthias R. Mehl[5]

[1]Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA. Email: kbollich@wustl.edu

[2]Department of Philosophy and Philosophy–Neuroscience–Psychology Program, Washington University in St. Louis, St. Louis, MO, USA. Email: jdoris@wustl.edu

[3]Department of Psychology, University of California, Davis, CA, USA. Email: svazire@ucdavis.edu

[4]Department of Psychiatry, University of Arizona, Tucson, AZ, USA. Email: craison@email.arizona.edu

[5]Department of Psychology, University of Arizona, Tucson, AZ, USA. Email: mehl@email.arizona.edu

*Corresponding author:
Kathryn L. Bollich
Campus Box 1125
Psychological & Brain Sciences
Washington University in St. Louis
One Brookings Drive
St. Louis, MO 63130-4899

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Abstract

Despite decades of interest in moral character, comparatively little is known about moral behavior in everyday life. This paper reports a novel method for assessing everyday moral behaviors using the Electronically Activated Recorder (EAR)—a digital audio-recorder that intermittently samples snippets of ambient sounds from people's environments—and examines the stability of these moral behaviors. In three samples (combined N = 186), participants wore an EAR over one or two weekends. Audio files were coded for everyday moral behaviors (e.g., showing sympathy, gratitude) and morally-neutral comparison language behaviors (e.g., use of prepositions, articles). Results indicate that stable individual differences in moral behavior can be systematically observed in daily life, and that their stability is comparable to the stability of neutral language behaviors.

KEYWORDS: moral character, moral behavior, Electronically Activated Recorder (EAR), temporal stability, personality, naturalistic observation, ambulatory assessment

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

"Living a moral, constructive life is defined by a weighted sum of countless individual, morally relevant behaviors enacted day in and day out (plus an occasional particularly self-defining moment)."

- Tangney, Stuewig, & Mashek (2007)

Morality has received a great deal of attention from psychologists in recent years. However, little of this work has examined moral behavior in naturalistic, "real-world" contexts. As such, the present study aims to establish a novel, reliable method for objectively and unobtrusively measuring moral behaviors that are observed in ordinary, everyday settings, and to use this method to examine the stability of individual differences in moral behaviors.

To place the current work into context, we highlight important gaps in the existing literature on morality. First, while classic social psychological research (e.g., Darley & Batson, 1973; Milgram, 1974) examined overt behavior, modern research has largely focused on moral cognition and emotion. Psychology has lately seen a surge of research on moral decision-making and the cognitive and emotional factors that influence moral judgments (Graham, Meindl, & Beall, 2012; c.f., Aquino & Freeman, 2009; Schwitzgebel, 2009), but, little contemporary work has examined overt moral behaviors, especially frequent, everyday moral acts (as opposed to exceptional moral acts).

To the extent that moral behavior has been studied, the research relies heavily on self-reported and laboratory-based measures. This is appropriate for research on moral identity, values, and judgments, but is problematic for studying moral behavior. People, on average, view themselves in a positive light (Alicke & Sedikides, 2009) and are especially likely to have distorted self-views for traits and behaviors that are highly evaluative (i.e., positively or negatively valenced; Vazire, 2010). Moral behaviors are arguably among the most evaluative

ACCEPTED MANUSCRIPT

behaviors (Goodwin, Piazza, & Rozin, 2013; Wojciszke, Bazinska, & Jaworski, 1998), which raises concerns about the accuracy of self-reports. Thus, although both self-views and behaviors are important to study and understand, self-reports of behavior are an inadequate substitute for measuring actual moral behavior (Graham, 2014).

At the same time, while studies *have* directly assessed moral behavior, these have mostly taken place in staged laboratory environments (e.g., Batson, Kobrynowicz, Dinnerstein, Kampf, & Wilson, 1997; Zhong, Bohns, & Gino, 2010; cf., Bateson, Nettle, & Roberts, 2006; Schwitzgebel, 2009). This methodology is insufficient for examining individual differences in moral behavior because people's laboratory behavior may not adequately reflect everyday behavior (Graham, 2014). Recent work has begun to explore morality in more natural contexts (Hofmann, Wisneski, Brandt, & Skitka, 2014), but this work frequently relies on self-reports of behaviors. To begin developing a more complete understanding of everyday moral functioning, the present study seeks to establish a reliable method for objectively observing moral behaviors outside the laboratory.

The existing literature emphasizes the variability of morality (Graham et al., 2012; Hartshorne & May, 1928), and how even subtle situational manipulations influence moral actions (Blanken, van de Ven, & Zeelenberg, 2015; Darley & Batson, 1973; Doris, 2002). Although this emphasis has sparked important research, relatively little of this work directly addresses the stability of individual differences in moral behavior. Where individual differences in morality have been examined, the focus has usually been on differences in moral perceptions and values (e.g., moral foundations; Graham, Nosek, Haidt, Iyer, Koleva, & Ditto, 2011), rather than actual moral behavior. To fill this gap, we examine the temporal stability of individual differences in actual, naturalistically observed, moral behaviors.

ACCEPTED MANUSCRIPT

Present Study

We present a method for objectively measuring everyday moral behaviors and examine the degree to which individual differences in these behaviors are stable across context and time. We use repeated observations in natural contexts to examine the consistency of moral behaviors—that is, whether people who act in more morally desirable ways than others at one time are also likely do so at another time. Our goal is to provide evidence for the viability of a new naturalistic method for studying actual, everyday moral behavior, as well as evidence about the degree to which individual differences in moral behavior are stable.

Our method employed the Electronically Activated Recorder, or EAR, a pocket-sized, wearable device that intermittently records short "sound-bites" of the wearer's audible environment, allowing researchers to unobtrusively capture ambient sounds from people's moment-to-moment lives (Mehl, Robbins, & große Deters, 2012). This method allows us to objectively assess actual behaviors, addressing calls to reemphasize the study of behavior in personality and social psychology (Baumeister, Vohs, & Funder, 2007; Furr, 2009). Although we are not able to assess all moral behaviors with this method, such as grand acts of heroism and self-sacrifice, this method does allow us to measure what is perhaps the most common form of morality (Hofmann et al., 2014): everyday, moral behaviors with a prosocial (or anti-social) focus. Additionally, the EAR enables the collection of representative samples from the full spectrum of participants' daily lives over several days, maximizing the generalizability and ecological validity of research findings (Brunswik, 1956) and allowing us to capture patterns of behavior that are more likely than single instances to reflect individual differences in moral personality.

ACCEPTED MANUSCRIPT

Following previous work on the stability of personality and self-reported moral constructs, we predicted that individual differences in moral behavior would be relatively stable over time (exhibiting moderate effect sizes of r = .30 to .50). This prediction is based on the test-retest stability of personality traits such as agreeableness and conscientiousness (Fleeson, 2001; Roberts & DelVecchio, 2000), which are related to behaving morally (Cohen, Panter, Turan, Morse, & Kim, 2014; Matsuba & Walker, 2004). Furthermore, studies examining explicitly moral constructs have shown that the rank-order stability of individual differences in moral judgments is relatively high (Bollich, Hill, Harms, & Jackson, 2015; Graham et al., 2011). Although most of this work relies on self- or other-reports of traits rather than observed behavior, it nevertheless provides grounds for predicting stable individual differences in moral behavior. The methodology of the present study enables us to directly test this prediction.

Method

Participants

We report how we determined our sample sizes, all data exclusions, and all relevant measures in the study. We used data from three samples, for a total of 186 participants[1]. Sample 1 consisted of 11 rheumatoid arthritis patients (11 women; M_{age} = 56.38, SD_{age} = 13.32; for more information on this sample, see Robbins, Mehl, Holleran, & Kasle, 2011). Sample 2 consisted of 73 adults participating in a randomized controlled trial of the effects of a meditation intervention on healthy adults (47 women, 26 men; M_{age} = 32.16, SD_{age} = 7.99; Raison, 2014). Sample 3 included 102 adults consisting of 52 women with breast cancer undergoing adjuvant cancer treatment (M_{age} = 56.16, SD_{age} = 13.95) and their co-habitating partners (7 women, 43 men; M_{age}

[1] Samples 1 and 2 were used in previous work (Robbins, Focella, Kasle, López, Weihs, & Mehl, 2011; Robbins, Lopez, Weihs, & Mehl, 2014; Robbins, Mehl, Holleran, & Kasle, 2011). However, the present analyses do not overlap with those in previous publications. A broad overview of the project was summarized in Mehl, Bollich, Vazire, & Doris (2015).

ACCEPTED MANUSCRIPT

= 59.41, SD_{age} = 14.61; for more information on this sample, see Robbins, Lopez, Weihs, & Mehl, 2014). Sample sizes were determined by the availability of resources and preexisting data. These sample sizes provided 80% power to detect effect sizes of at least *r* = .69 for Sample 1, at least *r* = .32 for Sample 2, and at least *r* = .28 for Sample 3. (Data from the three samples can be accessed at https://osf.io/xpqhw/.) For some analyses, the three samples were combined into one dataset—we note below when this is the case.

Participant Procedure

All participants wore the Electronically Activated Recorder (EAR; Mehl et al., 2012), a small electronic recording device that turns on intermittently and records sound-bites from participants' daily lives over the course of the study. The EAR consists of a mobile device (HP iPAQ 110 or Apple iPod touch) and a recording application (a software for the HP device and an iTunes App for the iPod touch). Participants in Sample 1 wore the EAR on two weekends four weeks apart, and it recorded 50 s every 18 min (average number of valid files with audible speech = 101, *SD* = 43). Participants in Sample 2 wore the EAR on two weekends about 10 weeks apart, and it recorded either 50 s every 9 min or 30 s every 12 min (average number of valid files with audible speech = 137, *SD* = 58). Participants in Sample 3 wore the EAR on one weekend, and it recorded 50 s every 9 min (average number of valid files with audible speech = 78, *SD* = 36). In all three samples, participants were informed their files would be coded for a broad set of daily behaviors, but moral behaviors were not specifically mentioned.

EAR Coding and Transcribing

In total, 19,063 EAR files containing audible speech were coded and transcribed by trained research assistants. For Samples 1 and 2, each research assistant who coded the first weekend files for a given participant also coded that same participant's files from the second

ACCEPTED MANUSCRIPT

weekend. For all files in which participants were talking, coders coded each file for a set of positive and negative moral behaviors.

Naturalistically observable moral behaviors. The first step in this research was determining what moral behaviors could be coded from the EAR data. We sought to identify behaviors that are both acoustically detectable and occur with some regularity in everyday life. This necessarily excluded rare behaviors (e.g., rescuing a drowning child) and emphasized interpersonal behaviors (because they are more common and more likely to be acoustically detectable), while excluding solitary behaviors without a readily audible component (e.g., cheating on a test). In addition, because we had only the auditory channel and not other channels of information (e.g., visual) available, we were limited to verbal behaviors (i.e., words) as opposed to physical acts (e.g., gestures). However, many everyday moral acts are expressed in words (e.g., apologizing, criticizing), and so although more channels of information would clearly have provided added value, we arguably had the most important single channel (auditory) for detecting a broad range of everyday moral behaviors.

We sought to cover a range of everyday moral behaviors whose presence (e.g., showing sympathy) or absence (e.g., acting condescending) is indicative of moral conduct. Although moral conduct includes more than just prosocial (or antisocial) actions, recent research shows that prosocial behaviors (i.e., behaviors related to harm or care) are the most commonly self-reported type of moral behavior in daily life (Hofmann et al., 2014), so, moral behavior was defined as behavior with a prosocial (or antisocial) focus. Using these three criteria (audibly detectable, occurring regularly, and prosocial), we selected 14 categories of moral behavior (e.g., showing affection, showing gratitude, praising or complimenting, acting condescending or arrogant, criticizing others) that the authors determined would be sufficiently audible and

ACCEPTED MANUSCRIPT

common, and which pilot testing suggested could be reliably assessed. Table 1 provides a complete list of the 14 behavioral categories and examples from actual EAR recordings (more detailed information on considerations around the development of an EAR coding system and the implementation of EAR behavior coding are provided in the online Supplemental Material).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 1. Everyday Positive and Negative Moral Behaviors and Examples from Actual EAR Sound Files.

Everyday Moral Behaviors	Examples (adapted from EAR sound files)
Positive Behaviors	
Showing affection	*"I love you; I really do."*
Showing gratitude	*"Thank you so much! I really appreciate it. You helped me a lot!"*
Offering praise, making compliments	*"You're learning a lot. You are doing great!"*
Showing sympathy, empathy, concern	*"I don't know why she did such hurtful things to you."*
Offering help, support	*"Can I carry that for you?"*
Apologizing	*"I am so sorry; I didn't mean to make you look bad."*
Expressing hope, showing optimism	*"I really want to be proactive about that. It'll be great."*
Negative Behaviors	
Being sarcastic	*"You want my lemonade?? Sure! After you drank the whole thing!"*
Bragging	*"My dad's car collection is so large he had to build a new garage."*
Being condescending or arrogant	*"How did you get so retarded? Oh, yes. I know what it means."*
Complaining or whining	*"You make it so difficult!"*
Criticizing	*"Then get off your butt and fix them lunch."*
Blaming	*"I came to see you and you were gone! You made me wait an hour!"*
Expressing pessimism	*"I doubt this medication will help."*

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Coding procedure. All sound files in which the participant was talking were coded by three independent coders for the presence or absence of all 14 moral behaviors. Any one file could contain more than one kind of moral behavior (e.g., a participant might apologize and show gratitude in one file). This behavior counting approach to EAR coding complements existing behavioral observation studies that are based on validated behavior rating systems (Funder, Furr, & Colvin, 2000), and yields data that are based on non-arbitrary, and intuitively interpretable metrics (i.e., the number of times a behavior was displayed or the percentage of interactions in which a behavior was present), thereby facilitating effect size calibration and interpretation (Blanton & Jaccard, 2006).

Naturalistically observable neutral behaviors. How much temporal stability is necessary to say that a behavior is stable? To provide a benchmark against which to compare the stability of individual differences in moral behaviors, we computed the stability of individual differences for a set of neutral behaviors. To select these neutral behaviors, "morally empty" language variables were chosen that matched the moral behaviors in terms of base rate. The categories we chose were: articles (e.g., a, the), prepositions (e.g., of, between), adverbs (e.g., around, here) and references to space (e.g., above, near), time (e.g., early, bye), and numbers (e.g., first, five).

These language variables provide a particularly strong test for several reasons. First, language can be coded objectively and reliably from EAR files (Mehl & Pennebaker, 2003), which means that stability estimates of language will not be attenuated due to unreliability. The selected language variables are maximally evaluatively neutral—that is, they have minimal inherent positive or negative connotation (e.g., article and preposition use), and thus are (largely) theoretically and empirically independent of the examined moral behaviors examined (average |*r*|

ACCEPTED MANUSCRIPT

= .11). Finally, these neutral language behaviors are not likely to be subject to strong self-presentational effects (i.e., participants are not likely to intentionally vary their use of prepositions from one context to another for self-presentational reasons). For these reasons, we expected individual differences in the neutral language behaviors to be reliable and stable, and thus a high benchmark for judging the stability of moral behaviors.

Transcribing procedure. Trained coders transcribed all of participants' utterances contained in their EAR files. Specifically, one coder transcribed the conversations while coding other behaviors, and subsequent coders proofread transcripts while coding for other behaviors.

Data Preparation

In all analyses, we only used EAR files with audible speech by the participant, as these are the only files in which participants could be (audibly) performing the moral and language behaviors of interest to the present study.

Inter-rater reliability of moral behavior coding. We calculated the inter-rater reliability for each of the moral behaviors separately in each sample and with all three samples combined[2]. For each participant, we averaged each coder's ratings of each moral behavior across all of a participant's EAR files. This provided us three average ratings per participant (one from each coder) for each of the moral behaviors we examined. Table 2 (columns 4, 7, 10, and 13) shows the inter-coder reliability (ICC[1,3]) for each moral behavior.

[2] The combining of all three samples does not take into account some dependence among participants in Sample 3. Ignoring this non-independence (resulting from within-couple similarity) should leave our effect estimates unbiased but might result in biased standard errors (and confidence intervals) and an overestimation of the effective degrees of freedom (Kenny et al., 2006). Given that the combined sample size is large (N = 186) and we tested for zero-order effects, using slightly biased standard errors and significance tests seemed preferable over (randomly) excluding data from one member of each dyad in Sample 3.

ACCEPTED MANUSCRIPT

Table 2. Moral Behaviors and Neutral Language Behaviors: Means, Standard Deviations, and Inter-Coder Reliabilities

	Sample 1 N = 11			Sample 2 N = 73			Sample 3 N = 102			All Samples N = 186		
EAR Variable	*M*	*SD*	*ICC (1,3)*	*M*	*SD*	*ICC (1,3)*	*M*	*SD*	*ICC (1,3)*	*M*	*SD*	*ICC (1,3)*
Moral Behaviors												
Affection	3.62	3.79	.95	2.80	3.63	.89	5.28	5.77	.97	4.21	5.05	.95
Gratitude	2.91	1.82	.84	2.98	2.74	.93	3.39	2.97	.94	3.20	2.82	.93
Praise / Compliment	3.10	1.81	.35	3.02	2.64	.70	4.46	3.34	.65	3.81	3.08	.67
Sympathy / Concern	2.54	2.28	.59	3.85	3.39	.28	8.00	4.67	.73	6.05	4.62	.69
Help / Support	4.73	3.70	.16	11.89	10.49	.27	20.57	7.52	.45	16.22	10.00	.51
Apologize	0.99	0.90	.82	2.53	6.05	.98	1.06	1.49	.96	1.63	4.00	.98
Hope / Optimism	1.85	1.83	.04	2.17	2.74	.72	1.83	1.75	.48	1.96	2.19	.61
Sarcasm	0.75	0.53	-.51	0.84	0.94	.41	0.45	0.72	.36	0.62	0.83	.39
Brag	0.57	0.72	.59	0.40	0.62	.53	0.13	0.34	.53	0.26	0.51	.57
Condescending / Arrogant	1.11	1.23	.54	0.74	1.33	.38	0.39	0.75	.25	0.57	1.06	.38
Complain / Whine	4.73	4.35	.80	5.77	5.15	.69	9.60	6.77	.72	7.81	6.35	.74
Criticism	3.72	3.10	.74	3.95	3.48	.71	5.10	4.56	.80	4.57	4.11	.78
Blame	0.52	0.66	.56	0.51	0.72	.37	0.96	1.22	.50	0.76	1.04	.49
Pessimism	0.95	1.03	.36	0.80	1.13	.44	0.86	1.13	-.15	0.84	1.12	.10
Average	*2.29*	*1.98*	*.49*	*3.02*	*3.22*	*.59*	*4.43*	*3.07*	*.59*	*3.75*	*3.34*	*.63*
Neutral Language Behaviors												
Article	4.32	0.70	-	3.80	0.77	-	4.40	0.70	-	4.16	0.78	-
Preposition	9.71	1.38	-	8.21	1.52	-	9.00	1.19	-	8.73	1.40	-
Adverb	6.68	0.46	-	5.45	1.03	-	5.78	1.05	-	5.71	1.05	-
Space	5.93	0.68	-	4.75	0.90	-	5.47	1.01	-	5.21	1.03	-
Time	4.53	0.79	-	4.13	1.02	-	4.56	0.93	-	4.39	0.98	-
Number	1.11	0.50	-	1.25	0.68	-	1.62	0.83	-	1.45	0.78	-
Average	*5.38*	*0.75*	-	*4.60*	*0.99*	-	*5.14*	*0.95*	-	*4.94*	*1.00*	-

Note. There are no ICCs for language categories because these were coded by a computer program (LIWC). Means are percentages of files with audible speech.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

EAR transcript preparation and word category selection. Verbatim EAR transcripts were analyzed using the Linguistic Inquiry and Word Count text analysis program (LIWC; Pennebaker, Francis, & Booth, 2007). This program analyzed participants' word use and calculated the percentage of participants' total words spoken in which they used particular categories of words (e.g., prepositions). We selected word categories that (a) are evaluatively neutral (i.e., have no or minimal positive or negative connotation), and (b) had relatively similar base rates to the moral behaviors (using base rates from all samples combined; Table 2, column 11). These categories were: articles, prepositions, adverbs, space, time, and numbers.

Stability Analyses

We took two approaches to measuring the stability of moral behaviors and language use. First we assessed *rank-order stability*, which we were able to test using participants who wore the EAR on two separate weekends (i.e., Samples 1 and 2). We averaged each moral and language behavior for the first weekend and the second weekend separately. This gave each participant two composite scores for each of the behaviors, which were then correlated with each other to obtain a measure of rank-order stability.

We also considered an alternative form of temporal stability: *momentary stability*. Using a similar method to Epstein's early work on trait and behavior stability (1979, 1980), we grouped EAR files by odd-numbered and even-numbered files—that is, a person's first sound file was odd, her second sound file was even, and so on. We then averaged behaviors within the odd files and within the even files. This gave each participant two composite scores for each of the moral and language behaviors, which were then correlated with each other in order to measure momentary stability, which we calculated for Samples 1-3. Because Sample 3 was comprised of romantic partners, we accounted for the dependency between couple members by using the Actor

ACCEPTED MANUSCRIPT

Partner Interdependence Model (APIM; Kenny, Kashy, & Cook, 2006). Parameter estimates from structural equation models were constrained across partners after establishing that there were no differences between breast cancer patients and their partners on the behaviors examined. Effect estimates were standardized for interpretability and comparison with estimates from Samples 1 and 2.

An important feature of these two methods for assessing stability is that we are aggregating across multiple observations (e.g., each of the odd vs. even and first weekend vs. second weekend aggregates contains numerous EAR codings per participant). By aggregating behaviors spread over several days, we are able to reduce measurement error and improve the reliability of the measures of behavior (Epstein, 1979). Including multiple instances of behaviors also increases the likelihood that we are capturing a representative sample of each participant's situations. As a result, these aggregated assessments capture patterns of behavior that are more likely to reflect individuals' typical levels of behavior compared to a single instance of behavior, often observed in a laboratory.

Results

Descriptive Statistics

The percentage of files in which participants displayed the moral behaviors can be found in Table 2, along with base rates for the matched neutral language behaviors. The moral behaviors were modestly correlated with each other (average within-sample $|r|$ = .26, range = .03 - .71), suggesting that the moral behaviors we assessed were diverse and captured non-overlapping variance in participants' moral acts. There were substantial individual differences in how often participants engaged in the moral behaviors we examined. For example, although on average people expressed gratitude in only 3.2% of their conversations, one person expressed

ACCEPTED MANUSCRIPT

gratitude during 17.5% of her conversations, whereas 16 people never expressed gratitude in any of their EAR recordings. In addition, although on average people only criticized others in 4.6% of their conversations, one person criticized others in 22.2% of her conversations, whereas 10 people never criticized others in any of their EAR recordings.

Rank-Order Stability of Moral Behaviors and Language Use (Samples 1 and 2)

First, we examined whether some people regularly behave more morally than others. To test this, we examined the rank-order stability of behavior by correlating the aggregate of behavior over the first weekend with the aggregate of behavior over the second weekend for each behavior (using Samples 1 and 2 only). Individual differences in moral behavior were moderately stable (average r = .47 over 4 weeks in Sample 1 and .52 over 10 weeks in Sample 2; Table 3). This was comparable to the stability of individual differences in the neutral language behaviors (average r = .26 over 4 weeks in Sample 1 and .45 over 10 weeks in Sample 2; Table 3). Taken together, these results show that individual differences in moral behavior are relatively stable over time.

Momentary Stability of Moral Behaviors and Language Use (Samples 1-3)

Next, we examined the momentary stability of moral behaviors and neutral language behaviors. In all three samples, we created an aggregate of odd files and another aggregate of even files and then correlated the two aggregates separately for each sample. For Sample 3, we used the Actor Partner Independence Model (APIM; Kenny et al., 2006) to control for the dependency between couple members.[3] Moral behaviors evidenced moderate to strong momentary stability (Sample 1 average r = .42, Sample 2 average r = .71, Sample 3 average standardized b = .37; Table 4), and this was comparable to the momentary stability of neutral language behaviors (Sample 1 average r = .32, Sample 2 average r = .66, Sample 3 average

[3] This analytic method excluded two participants who did not have partner data.

ACCEPTED MANUSCRIPT

standardized b = .28; Table 4). Together, these findings show that a person's typical level of engaging in moral behavior is a reliable, stable characteristic.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 3. Rank-Order Stability of Everyday Moral Behaviors and Neutral Language Behaviors

	Sample 1 N = 11			Sample 2 N = 73		
EAR Variable	*r*	*CIs [95%]*	*p*	*r*	*CIs [95%]*	*p*
Moral Behaviors						
Affection	.64*	[.06; .90]	.035	.58*	[.41; .72]	<.001
Gratitude	.17	[-.48; .70]	>.250	.05	[-.18; .27]	>.250
Praise / Compliment	-.37	[-.79; .30]	>.250	.30*	[.08; .50]	.010
Sympathy / Concern	.67*	[.12; .91]	.023	.55*	[.37; .69]	<.001
Help / Support	.47	[-.18; .84]	.143	.83*	[.74; .89]	<.001
Apologize	.35	[-.31; .79]	>.250	.84*	[.76; .90]	<.001
Hope / Optimism	.76*	[.29; .93]	.007	.58*	[.40; .71]	<.001
Sarcasm	.03	[-.58; .62]	>.250	.25*	[.02; .46]	.031
Brag	.24	[-.42; .74]	>.250	.01	[-.22; .24]	>.250
Condescending / Arrogant	.75*	[.27; .93]	.008	.51*	[.31; .66]	<.001
Complain / Whine	.78*	[.33; .94]	.005	.78*	[.67; .86]	<.001
Criticism	.70*	[.17; .92]	.017	.58*	[.41; .72]	<.001
Blame	.08	[-.55; .65]	>.250	.34*	[.12; .53]	.003
Pessimism	.63*	[.05; .89]	.037	.47*	[.27; .63]	<.001
Average	*.47*	-	-	*.52*	-	-
Neutral Language Behaviors						
Article	.28	[-.38; .75]	>.250	.35*	[.13; .54]	.002
Preposition	.47	[-.18; .84]	.141	.51*	[.32; .66]	<.001
Adverb	-.22	[-.72; .44]	>.250	.36*	[.14; .54]	.002
Space	-.13	[-.68; .51]	>.250	.32*	[.10; .52]	.005
Time	.47	[-.18; .83]	.147	.53*	[.34; .67]	<.001
Number	.56	[-.05; .87]	.070	.57*	[.39; .70]	<.001
Average	*.26*	-	-	*.45*	-	-

Note. * $p < .05$, two-tailed. Rank-order stability is the correlation between the aggregate of the first weekend and the aggregate of the second weekend (approximately 4 weeks [Sample 1] and 10 weeks [Sample 2] apart).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 4. Momentary Stability of Everyday Moral Behaviors and Neutral Language Behaviors

	Sample 1 N = 11			Sample 2 N = 73			Sample 3 N = 100		
EAR Variable	*r*	*CIs [95%]*	*p*	*r*	*CIs [95%]*	*p*	*stand. b*	*CIs [95%]*	*p*
Moral Behaviors									
Affection	.72*	[.28; .91]	.005	.82*	[.73; .89]	<.001	.71*	[.55; .86]	<.001
Gratitude	.21	[-.38; .68]	>.250	.14	[-.09; .36]	.229	.27*	[.09; .46]	.004
Praise / Compliment	.21	[-.38; .68]	>.250	.53*	[.34; .68]	<.001	.31*	[.10; .52]	.003
Sympathy / Concern	.56*	[.02; .85]	.044	.80*	[.69; .87]	<.001	.51*	[.34; .69]	<.001
Help / Support	.62*	[.10; .87]	.024	.90*	[.85; .94]	<.001	.47*	[.33; .61]	<.001
Apologize	-.10	[-.62; .48]	>.250	.89*	[.83; .93]	<.001	.29*	[.11; .48]	.002
Hope / Optimism	.68*	[.20; .90]	.011	.77*	[.66; .85]	<.001	.12	[-.06; .31]	.190
Sarcasm	.16	[-.43; .65]	>.250	.36*	[.14; .54]	.002	.01	[-.19; .21]	>.250
Brag	.20	[-.40; .68]	>.250	.32*	[.10; .51]	.005	.02	[-.17; .21]	>.250
Condescending / Arrogant	.60*	[.08; .87]	.029	.83*	[.74; .89]	<.001	.50*	[.33; .67]	<.001
Complain / Whine	.70*	[.24; .90]	.008	.86*	[.78; .91]	<.001	.70*	[.57; .83]	<.001
Criticism	.49	[-.09; .82]	.091	.73*	[.61; .83]	<.001	.58*	[.43; .73]	<.001
Blame	-.04	[-.58; .52]	>.250	.52*	[.33; .67]	<.001	.15	[-.04; .35]	.120
Pessimism	.38	[-.22; .77]	.206	.61*	[.45; .74]	<.001	.22*	[.08; .37]	.003
Average	*.42*	-	-	*.71*	-	-	*.37*	-	-
Neutral Language Behaviors									
Article	-.13	[-.68; .51]	>.250	.66*	[.50; .77]	<.001	.28*	[.11; .47]	<.001
Preposition	.51	[-.13; .85]	.110	.85*	[.78; .91]	<.001	.32*	[.15; .49]	<.001
Adverb	-.12	[-.67; .52]	>.250	.64*	[.48; 76]	<.001	.46*	[.29; .64]	<.001
Space	.28	[-.39; .75]	>.250	.57*	[.39; .71]	<.001	.13	[-.07; .33]	.210
Time	.25	[-.41; .74]	>.250	.64*	[.48; .76]	<.001	.12	[-.06; .30]	.195
Number	.82*	[.44; .95]	.001	.51*	[.32; .66]	<.001	.32*	[.15; .49]	<.001
Average	*.32*	-	-	*.66*	-	-	*.28*	-	-

Note. * $p < .05$, two-tailed. Momentary stability is the relationship between the aggregate of odd files and the aggregate of even files. Odd and even files were separately aggregated in each sample and then correlated. In Sample 3, momentary stability was calculated using APIM to account for the dependency of partners (standardized effect estimates are provided).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Discussion

The present study establishes a novel method for naturalistically assessing everyday moral behaviors, and provides evidence that there are substantially stable individual differences in these moral behaviors. Indeed, individual differences in moral behavior were at least as stable as individual differences in neutral language behaviors. This is impressive—we expected neutral language behaviors to be highly reliable and stable (because they can be measured without coder error and because they are not subject to self-presentational concerns), and thus considered them a high benchmark for gauging the stability of moral behaviors.

These findings present important evidence that socially significant moral behaviors can be reliably observed in daily life using the Electronically Activated Recorder (EAR). Given the potential biases in both self- and peer-reports of morality, using the EAR provides a complementary way to assess morality that sidesteps these limitations, and could ultimately be used to examine the accuracy of self- and peer-reports of morality. By bringing the study of morality out of the lab and into the real world where moral behaviors naturally occur, this method opens up the study of moral character to a variety of questions that will deepen our scientific understanding of the complexities of morality. For instance, it is possible to conduct a study specifically designed to capture moral behaviors as they occur across different settings (e.g., home, work, and social contexts), allowing researchers to directly examine the cross-situational consistency of moral behaviors (Bleidorn & Denissen, 2015).

As with any study, there are limitations that deserve attention and provide ideas for future research. The measure of rank-order stability we use spans a short time period (4 or 10 weeks), and future work should examine the stability of moral behaviors over longer periods. Additionally, two of the samples include patient populations (i.e., people with rheumatoid

ACCEPTED MANUSCRIPT

arthritis [Sample 1] and breast cancer [Sample 3]), and thus it is important to conduct additional work exploring other ages, health groups, and cultures. Nevertheless, it is worthwhile to note that these samples increase the generalizability of these findings across diverse ages and groups compared to the typical sample of college students. Furthermore, given that individuals in these three samples were dealing with the stresses of cancer, coping with chronic illness, or involved in a meditation intervention, the substantial stability in moral behavior we observed may actually be less than what might be observed in other populations that are experiencing less extraordinary circumstances (i.e., not undergoing personal upheavals or interventions).

While providing important evidence of temporal stability, the present study does not directly address the consistency of moral behaviors across diverse situations. Our results suggest there is stability from day to day, and as Epstein (1980) points out, because no two situations can be exactly the same, these findings indirectly suggest the presence of some cross-situational consistency. However, momentary stability does not strictly test consistency across situations and future research that simultaneously measures situations and daily moral behaviors will provide a more formal test of the stability of moral behaviors across contexts. Moreover, we cannot rule out the possibility that the stability of individual differences in moral behavior is due to the stability of individual differences in situations. That is, if there are individual differences in the situations people consistently find themselves in, this could be driving individual differences in moral behavior. However, the stability of individual differences in situations could itself be driven by personality differences (i.e., situation selection effects), so disentangling these processes will require extensive repeated assessments of situations and behavior over time.

Although the EAR offers a unique look at everyday, naturally-occurring moral behaviors, it also limits the type of moral behaviors that can be assessed. For example, we could not assess

ACCEPTED MANUSCRIPT

inaudible behaviors, such as cheating or dishonesty, or uncommon behaviors, such as acts of bravery or heroism. Nor could we fully assess the context in which behaviors occur because the sound bites are brief (<1 min), making it difficult to assess more complex moral behaviors. Like other behavioral research, the EAR does not allow confident attributions of mental states: the data do not allow us to speak to whether a participant had morally desirable or undesirable motivations or intentions. Finally, as evidenced by some lower intercoder reliabilities (Table 2), some moral behaviors are more difficult to code from the EAR. It is worth noting that behaviors with lower intercoder reliabilities also had lower base rates. However, we do not know if these base rates are specific to our samples, and encourage future research to continue assessing these behaviors (see Supplemental Material for recommendations for coding in future samples).

Despite these limitations, the present findings make important contributions to our understanding of individual differences in moral behavior. In addition, the use of the EAR to study moral behavior is an important advance in the study and measurement of moral behavior. Future research should examine how individual differences in moral behavior are related to self- and other-perceptions of morality, as well as moral judgments, emotions, and intentions. Together, these approaches will help us capture a more complete picture of morality as it is manifested in everyday life.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

References

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology, 20*, 1-48. doi:10.1080/10463280802613866

Aquino, K., & Freeman, D. (2009). Moral identity in business situations: A social-cognitive framework for understanding moral functioning. In D. Narvaez, D. K. Lapsley, D. Narvaez, D. K. Lapsley (Eds.), *Personality, identity, and character: Explorations in moral psychology* (pp. 375-395). New York, NY, US: Cambridge University Press. doi:10.1017/CBO9780511627125.018

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters, 2*(3), 412-414.

Batson, C., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology, 72*(6), 1335-1348. doi:10.1037/0022-3514.72.6.1335

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?. *Perspectives on Psychological Science, 2*(4), 396-403. doi:10.1111/j.1745-6916.2007.00051.x

Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin, 41*(4), 540-558. doi:10.1177/0146167215572134

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*(1), 27-41. doi:10.1037/0003-066X.61.1.27

Bleidorn, W., & Denissen, J. A. (2015). Virtues in action—The new look of character traits.

ACCEPTED MANUSCRIPT

British Journal of Psychology, *106*(4), 700-723. doi:10.1111/bjop.12117

Bollich, K. L., Hill, P. L., Harms, P. D., & Jackson, J. J. (2015). When friends' and society's expectations collide: A longitudinal study of moral decision-making and personality across college. Manuscript under review.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments (2nd ed.)*. Berkeley, CA, US: University of California Press.

Cohen, T. R., Panter, A. T., Turan, N., Morse, L., & Kim, Y. (2014). Moral character in the workplace. *Journal of Personality and Social Psychology*, *107*(5), 943-963. doi:10.1037/a0037245

Darley, J. M., & Batson, C. (1973). 'From Jerusalem to Jericho': A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, *27*(1), 100-108. doi:10.1037/h0034449

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge University Press.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*(7), 1097-1126. doi:10.1037/0022-3514.37.7.1097

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist*, *35*(9), 790-806. doi:10.1037/0003-066X.35.9.790

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011-1027. doi:10.1037/0022-3514.80.6.1011

Funder, D. C., Furr, R., & Colvin, C. (2000). The Riverside Behavioral Q-sort: A tool for the

ACCEPTED MANUSCRIPT

description of social behavior. *Journal of Personality*, *68*(3), 451-489. doi:10.1111/1467-6494.00103

Furr, R. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality*, *23*(5), 369-401. doi:10.1002/per.724

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148-168. doi:10.1037/a0034726

Graham, J. (2014). Morality beyond the lab. *Science*, 345 (6202), 1242-1242. doi: 10.1126/science.1259500

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366-385. doi:10.1037/a0021847

Graham, J., Meindl, P., & Beall, E. (2012). Integrating the streams of morality research: The case of political ideology. *Current Directions in Psychological Science*, *21*(6), 373-377. doi:10.1177/0963721412456842

Hartshorne, H., & May, M. (1928). *Studies in the nature of character. Vol. 1. Studies in deceit*. New York, NY: MacMillan.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, 345 (6202), 1340-1343. doi: 10.1126/science.1251560

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. Guilford Press.

Matsuba, M., & Walker, L. J. (2004). Extraordinary moral commitment: Young adults involved in social organizations. *Journal of Personality*, *72*(2), 413-436. doi:10.1111/j.0022-3506.2004.00267.x

ACCEPTED MANUSCRIPT

Mehl, M. R., Bollich, K. L., Doris, J. M., & Vazire, S. (2015). Character and coherence: Testing the stability of naturalistically observed daily moral behavior. In C. Miller, A. Knobel, R. M. Furr, & W. Fleeson (Eds.), *Character: New Directions from Philosophy, Psychology, and Theology*. Oxford University Press.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology, 84*(4), 857-870. doi:10.1037/0022-3514.84.4.857

Mehl, M. R., Robbins, M. L., & Deters, F. G. (2012). Naturalistic observation of health-relevant social processes: The electronically activated recorder methodology in psychosomatics. *Psychosomatic Medicine, 74*(4), 410-417. doi:10.1097/PSY.0b013e3182545470

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2007). Linguistic inquiry and word count (LIWC): LIWC 2007 [Computer program]. Mahwah, NJ: Erlbaum.

Raison, C. L. (2014). The sounds of compassion: Testing how specific elements of meditation change daily life. ClinicalTrials.gov (ClinicalTrials.gov identifier: NCT01643369).

Robbins, M. L., Focella, E. S., Kasle, S., López, A. M., Weihs, K. L., & Mehl, M. R. (2011). Naturalistically observed swearing, emotional support, and depressive symptoms in women coping with illness. *Health Psychology*, *30*(6), 789-792. doi:10.1037/a0023431

Robbins, M. L., López, A. M., Weihs, K. L., & Mehl, M. R. (2014). Cancer conversations in context: Naturalistic observation of couples coping with breast cancer. *Journal of Family Psychology*, *28*(3), 380-390. doi:10.1037/a0036458

Robbins, M. L., Mehl, M. R., Holleran, S. E., & Kasle, S. (2011). Naturalistically observed sighing and depression in rheumatoid arthritis patients: A preliminary study. *Health Psychology*, *30*(1), 129-133. doi:10.1037/a0021558

ACCEPTED MANUSCRIPT

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*(1), 3-25. doi:10.1037/0033-2909.126.1.3

Schwitzgebel, E. (2009). Do ethicists steal more books?. *Philosophical Psychology*, *22*(6), 711-725. doi:10.1080/09515080903409952

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58,* 345-372. doi:10.1146/annurev.psych.56.091103.070145

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281-300. doi:10.1037/a0017908

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251-1263. doi:10.1177/01461672982412001

Zhong, C., Bohns, V. K., & Gino, F. (2010). Good lamps are the best police: Darkness increases dishonesty and self-interested behavior. *Psychological Science*, *21*(3), 311-314. doi:10.1177/0956797609360754

ACCEPTED MANUSCRIPT