Philosophical Psychology, 2014
http://dx.doi.org/10.1080/09515089.2014.893868

The Free-Will Intuitions Scale and the question of natural compatibilism

Oisín Deery, Taylor Davis and Jasmine Carey

Standard methods in experimental philosophy have sought to measure folk intuitions using experiments, but certain limitations are inherent in experimental methods. Accordingly, we have designed the Free-Will Intuitions Scale to empirically measure folk intuitions relevant to free-will debates using a different method. This method reveals what folk intuitions are like prior to participants' being put in forced-choice experiments. Our results suggest that a central debate in the experimental philosophy of free will—the "natural" compatibilism debate—is mistaken in assuming that folk intuitions are exclusively either compatibilist or incompatibilist. They also identify a number of important new issues in the empirical study of free-will intuitions.

Keywords: Compatibilism; Experimental Philosophy; Free Will

1. Introduction

Do people naturally think free will is compatible with determinism? Or are they instead natural incompatibilists? This has been a central debate in the experimental philosophy of free will, and the standard approach of those addressing this question has been to ask the folk for intuitive judgments in response to vignettes, where these judgments take the form of choices between conflicting philosophical positions. A limitation of this approach, which we call the Conflict Method, is that it is committed to a conception of intuitive judgments that is limited, and can be misleading. As a way of supplementing this method, we have developed a philosophical free-will scale, following the scale methodology used in social psychology. Our *Free-Will Intuitions Scale* (FWIS) provides access to psychological information that is inaccessible to the Conflict Method, revealing what folk intuitions

Oisín Deery is a GRIN Postdoctoral Research Fellow in philosophy at the Center for Research in Ethics, University of Montreal.
Taylor Davis is a Ph.D. student in philosophy at the University of British Columbia.
Jasmine Carey is a Ph.D. student in psychology at the University of British Columbia.
Correspondence to: Oisín Deery, University of Montreal—Center for Research in Ethics, 2910 Boul. Édouard-Montpetit, Montréal, Quebec H3T 1J7, Canada. Email: oisin@oisindeery.com

are like prior to participants' being placed in forced-choice experimental situations. We discuss here the impact of this information on the question of natural compatibilism.

Unlike previous free-will scales developed by psychologists, our scale is designed from the ground up with philosophical issues in mind. The process of developing a scale allows us to determine empirically when the folk will interpret philosophical statements in ways relevant to philosophical debates. These data are important, and are missed by the Conflict Method. We emphasize, however, that the FWIS is a supplement to the Conflict Method, not a replacement for it. The Conflict Method tells us *how* individuals resolve conflicts between competing intuitions, whereas the FWIS explains *why* people resolve conflicts of intuition in one way rather than another.

In section 2, we say more about the difference between the information gathered by the Conflict Method and that gathered by the scale method. In section 3, we explain the factor-analytic method that we employ, and the specific questions we hope to address by using this method. In section 4 we present our results, before turning to a general discussion in section 5.

2. Natural Compatibility and the Conflict Method

Some experimental philosophers have recently conjectured that folk intuitions are *naturally* incompatibilist (e.g., Nichols, 2004; Nichols & Knobe, 2007). Others deny this, presenting empirical results designed to show instead that people are natural compatibilists (e.g., Nahmias, Morris, Nadelhoffer, & Turner, 2005; Nahmias & Murray, 2011; compare Feltz, Cokely, & Nadelhoffer, 2009). The compatibility at issue is between free will and determinism, where determinism is the thesis that the facts of the past and the laws of nature together entail just one physically possible future.

Rather than directly addressing metaphysical questions, experimental philosophers focus on the *natural* compatibility question, which is a psychological question about the pre-theoretic intuitions that lead people to endorse claims about the compatibility of determinism and freedom. This question concerns whether people *begin* as compatibilists or incompatibilists, prior to their considering philosophical theories. If someone outside the context of philosophical theorizing assents to a statement about free will that is logically incompatible with determinism, then that person is, to that extent, a natural incompatibilist—whether or not she is explicitly aware of the incompatibility. Compatibilism and incompatibilism are *logically* incompatible, but that is no reason to assume that they are *psychologically* incompatible, and the question of natural compatibilism is above all a psychological question.

Answering this question still leaves open normative issues about how we *ought* to think about free will, since there may be good reasons why we *should* think differently than we do. Nonetheless, answers to psychological questions may inform our approach to normative theorizing. Indeed, questions of natural compatibility first took root in experimental philosophy because of claims made by both compatibilist and incompatibilist philosophers to the effect that, since the folk are natural (in)compatibilists, the burden of proof lay on the opposing camp to show that our natural position is in need of revision (e.g., Nahmias et al., 2005). One response to this challenge is to deny that people *are* natural (in)compatibilists in the first place. This requires doing empirical work to show what people's natural intuitions actually are, and much of the literature in experimental philosophy aims to do just this.

However, almost a decade after philosophers first began designing experiments to measure folk intuitions about (in)compatibilism, this issue remains unresolved, with published results supporting each side. This has led researchers to propose error theories for their opponents' results. In a number of papers, Nahmias and his collaborators attempt to explain away incompatibilist intuitions by hypothesizing that the folk misinterpret determinism in one of two ways, either as mechanism or as fatalism. On the other side, Nichols and Knobe (2007) offer an error theory for *compatibilist* intuitions, arguing that these judgments occur only when affective responses cloud the judgment of participants.[1] We think this empirical stalemate is due in part to a limitation in how experimental philosophers have been thinking about intuitions.[2]

Experimental philosophers treat intuitions as the *outcomes of decisions*, so we call these "decision" intuitions. Participants are first situated within a philosophical debate, usually by reading vignettes, and are asked to take sides in that debate, usually by making forced-choice responses: for instance, agents in a deterministic universe either are or are not morally responsible (Nichols & Knobe, 2007). Yet in many cases, it seems likely that participants will find *both* options intuitively compelling, though perhaps for different reasons, and to differing degrees. By requiring a decision between these "basic" intuitions, this experimental design elicits an internal conflict in the mind of the respondent, and response data only record the outcomes of such conflicts. Accordingly, we call this the Conflict Method.[3]

Note that basic intuitions are not a new type of decision intuition. Basic intuitions are, like decision intuitions, "aggregate" judgments formed from one's background beliefs, inferences, and other mental states; we don't suggest that they are basic tout court, as psychological entities. Rather, they are basic relative to the decision intuitions of the Conflict Method. They are complex psychological processes, but they comprise options that participants must choose between in Conflict Method studies. Basic intuitions are inputs to the decision process, while decision intuitions are outputs. What isn't recorded by the Conflict Method is information about this deliberative process itself—the motivational struggle between basic intuitions that *produces* a decision intuition. "The intuition" of the respondent is only the "winning" intuition; all information about "losing" intuitions is lost, and the results are treated as if no struggle had occurred. This makes it look as though all respondents took their answers to be obvious. The Conflict Method thus leaves out critical information about the target phenomena of experimental philosophy: psychological facts about philosophically relevant intuitions. As a result, data collected in this way are often misleading.[4]

Consider a schematic worst-case scenario. Out of 100 subjects, 50 people find option B mildly attractive, but option A is clearly better. The other 50 people, by contrast, find the decision very difficult, and make their choices without conviction or confidence. The internal conflict is intense, and respondents feel they are choosing something just because they have been asked to do so—they might as well have flipped a coin. Accordingly, the indecisive group splits evenly down the middle, with 25 people choosing A and 25 choosing B, and in the end 25 people choose B and 75 choose A. For the Conflict Method, "the intuition" of the folk is clearly represented by choice A.

From the point of view of psychological methodology, this is a disaster. Psychologically speaking, there is no difference between the 25 people who chose A for no good reason and the 25 people who chose B for no good reason. The important difference in this sample lies between the decisive group and the indecisive group, but the experimental design can't register this difference. The conclusion drawn is that the folk overwhelmingly prefer choice A, when fully half of the sample didn't prefer choice A at all.[5] We found something similar for free-will intuitions: people are often motivated to be *both* compatibilists *and* incompatibilists. We couldn't have discovered this just by running experiments asking individuals to *choose between* compatibilism and incompatibilism.[6]

Certainly, how people resolve conflicts of intuition is an important part of the empirical story, and it should be told. Indeed, our scale is intended to be administered *along with* studies employing the Conflict Method. By offering an alternative way of operationalizing intuitions, the FWIS identifies relationships *across* experimental studies, thus providing a guide to which experiments most need to be run, which should take priority over others, and so on. Of course, information missed by any Conflict study *could,* in principle, be accessed by another Conflict study with a different design.[7] Yet one would have to run a large number of carefully engineered Conflict studies to reproduce the data gathered by a single administration of the FWIS. The Conflict Method makes it difficult to examine relationships *between* basic intuitions and decision intuitions, whereas our method makes it easy.

The FWIS is modeled after the personality scales of social psychology (e.g., John & Srivastava, 1999). Of course, we are not the first to use scale methods to study the psychology of free will (e.g., Paulhus & Carey, 2011; Rakos, Laurene, Skala, & Slane, 2008; Stroessner & Green, 1990; Viney, Waldman, & Barchilon, 1982). Yet, existing free-will scales were designed by psychologists to address psychological concerns, and they are not ideal for philosophical purposes, since the items comprising them don't express coherent philosophical positions.[8] Using scale methods for the purpose of studying philosophical intuitions poses a distinct methodological challenge, since we must elicit judgments about philosophical views without asking participants to choose sides in a debate. In the next section, we show how the FWIS meets this challenge, and this marks yet another sense in which the scale provides access to information that is missed by the Conflict Method. The process of developing a scale provides empirical data justifying the assumption that participants' judgments are, in fact, judgments *of* the philosophical positions of interest. Since questions of correct interpretation have been a source of disagreement in recent experimental philosophy, these data bear on some existing findings. First, however, we describe how a philosophical free-will scale works, and what it shows.

3. Scale Methodology and Factor Analysis

We conducted four rounds of data collection in order to identify the kinds of statements that belong on a free-will scale—statements that are unlikely to be misinterpreted in ways that render them irrelevant to philosophical debates. Participants were asked to report the degree to which they agreed or disagreed, on a seven-point Likert scale, with items in a randomized list. Responses were analyzed using exploratory factor analysis to determine which items formed coherent factors, or patterns of agreement across distinct items.

According to this method of analysis, distinct factors appear when participants respond in similar ways to coherent groups of items. Each item is then given a score or "loading" for each factor, which describes the extent to which that item is responsible for the overall pattern represented by the factor. Since there isn't any reason for two people (much less two *hundred* people) to respond in the same way to items they interpret in different ways, the best explanation for the appearance of a factor is that participants are interpreting and responding to the items of that factor in the same way.

When two items load on the same factor, this indicates similarity in responses in two ways. First, the degree of similarity in responses *across multiple items* is what determines *which* factor(s) those items belong to. Second, the degree of similarity among responses *to a particular item* determines *how high or low* that item's loading is on a given factor.

As a result, when several items expressing the same philosophically interesting notion (e.g., determinism) load highly on the same factor, the best explanation for this is that participants share with one another a single conception of the position in question, which leads them to agree to statements of that position to similar degrees. When items that *don't* express the same philosophical notion load highly on the same factor, the best explanation for this pattern of responses is *also* that participants share with one another a single view that leads them to agree to all these statements to a similar extent. Thus, if the items comprising a factor don't express a coherent philosophical position, the best explanation for this is that participants are not recognizing the logical conflicts that threaten the coherence of that "position." Participants are conflating the statements in question, and misinterpreting them as statements of the same position. Items that don't load on a given factor are simply being interpreted by participants as having no relationship to the items that *do* load on that factor.

Using this method, we wanted to address a number of questions. First, we wanted to know whether people *naturally* distinguish compatibilist from incompatibilist statements of free will, when they consider these statements in isolation, and not in the context of a debate. As it turns out, they do. We found that compatibilism and incompatibilism emerged as distinct factors for two different notions of free will that we tested. Our results identified eight factors, which comprise distinct sub-scales on our final scale. Thus, our final scale (appendix A) consists of eight sub-scales, made up of four or five items each, examining the following eight positions relevant to free-will debates: ability-to-do-otherwise compatibilism; ability-to-do-otherwise incompatibilism; sourcehood compatibilism; sourcehood incompatibilism; moral responsibility; proximal-determinism; distal-determinism; and fatalism.[9]

The structure of the FWIS thereby addresses two different notions of free will taken from the philosophical literature: ability-to-do-otherwise freedom (ATDO) and sourcehood freedom (SH). Our items were designed to capture compatibilist and incompatibilist versions of each notion. For example, a typical compatibilist ATDO item read, "Audrey might have chosen to take the job in St. Louis instead of the job in Toronto, but only if she had wanted the job in St. Louis more."[10] The idea behind compatibilist claims that one "could have done otherwise" is that they are conditional claims about what one *would* have done *if* something prior to the action had been different.[11] By contrast, incompatibilists think that one "could have done otherwise" even holding fixed everything prior to one's action. Thus, a typical incompatibilist ATDO item read, "I could have bought a different dish soap than I actually bought, and I might have done so even if none of my preferences or desires had been different."[12]

Of course, this isn't the only notion of free will. Many philosophers prefer to think about free will in terms of an agent's being the *source* of her actions, where this *doesn't* require being able to do otherwise.[13] Such "sourcehood" (SH) views require instead that the action in question issue in the right way from one's reasons, values, desires, and so forth. Many compatibilists think that this is sufficient for an action to count as free.[14] Thus:

> As long as Hannah decides what to do on the basis of her own reasons, that's sufficient for her to be the ultimate source of her actions; in other words, that's enough for her actions to be "up to her."

By contrast, incompatibilists require more than this, and incompatibilist sourcehood views are best expressed as a denial of the compatibilist's sufficiency claim, thus:

> Even when Owen decides what to do on the basis of his own reasons, that's not enough for him to be the ultimate source of his actions; he must also have had the final say about how he responds to such reasons.[15]

For each way of conceptualizing free will (ATDO and SH), separate factors emerged for compatibilist and incompatibilist versions of the view. So, participants in our studies *did* naturally distinguish compatibilist from incompatibilist free will, along these two dimensions.

Second, we wanted to know whether participants would be *exclusively* either natural compatibilists or natural incompatibilists. The answer is: they aren't (as we explain in detail below). While compatibilism and incompatibilism are *logically* incompatible, our results show that these positions aren't *psychologically* incompatible. People are sometimes both natural compatibilists and natural incompatibilists, even about the same notion of free will.[16]

Third, establishing which items belong on the FWIS, and what the factors are, enables us to address recent controversies in experimental philosophy about how notions like determinism should be operationalized. In particular, we wanted to find out whether participants would confuse determinism with mechanism or fatalism, as Nahmias and his collaborators suggest. We discovered that they don't—at least not in the way that Nahmias and colleagues predict.

Finally, we discovered some intriguing new patterns in how people think about free will. For instance, it is an empirical question what notion of free will (if any) is naturally most closely related to intuitions about moral responsibility. The answer provided by the FWIS is: sourcehood freedom. Indeed, we found a clustering of intuitions around two very different sorts of reason people have for being concerned about free will, each of which interacted differently with different formulations of determinism. This suggests a clear divide among different reasons for interpreting determinism as a threat to free will, as we discuss later.

4. Data Collection

4.1. Method

Four samples of 250 subjects were recruited using Mechanical Turk. Here, details are given for the fourth dataset, which established the final version of the FWIS. Items were presented in seven-point Likert format with anchors of 1 (strongly disagree) to 7 (strongly agree). Individual items were randomly assigned to different pages, which were displayed by a computer program in a random order. Demographic questions were included, as well as questions about philosophical education.

4.2. Results

4.2.1. Exploratory factor analysis

Because we collected data in four different rounds, we performed an Exploratory Factor Analysis (EFA) on each dataset independently. The analyses were run in SPSS, using Maximum Likelihood estimation and Varimax rotation.

In the fourth round of data collection, in which the numbers of items were balanced across the relevant philosophical positions, our hypothesized factors were found, with items loading on the intended factors. In the second round, however, distinct factors emerged for proximal-determinism (Proximal-D) and distal-determinism (Distal-D), which we had not hypothesized.[17] ATDO-Compatibilism and ATDO-Incompatibilism were clearly distinct factors, with no items cross-loading between the factors; mean cross-loading was –.08. By contrast, while SH-Compatibilism and SH-Incompatibilism formed separate factors, they were not as clearly distinct, with an average cross-loading of .24. Yet the cross-loadings were not large enough to warrant combining the factors, and all cross-loadings were lower than the minimum loadings on the intended factor.

Over the course of the four rounds of data collection, we addressed many concerns regarding wording. This led to fluctuations in loadings, which ultimately determined which items would comprise the scale, but it had no effect on the overall factor structure. Between the second and third rounds, wordings were added to the Proximal-D items to assess the impact of mechanistic (neural) or psychological (thoughts, desires) causes of action. These wordings had no effect on the loadings for the Proximal-D factor, and both wordings are included in the final item set. In the final study, both first- and third-person determinism items were included, which had no effect on the loadings. We also attempted to standardize the complexity and length of the items across the sub-scales, to ensure that these considerations were not responsible for creating the factors. The final item set appears in appendix A. The factor loadings are shown in Table 1.

Internal consistency reliability, usually measured with Cronbach's alpha, is used as an approximation of how similar an individual's score would be if given the scale again. Alphas were strong for all sub-scales. Distal-D had the highest reliability, $\alpha = .94$, while SH-Incompatibilism was the least reliable, $\alpha = .72$, but all were within an acceptable range, with .70 usually considered adequate and .90 being excellent. Strong reliability indicated that the items within each factor are strongly related to each other, and should remain stable across administrations of the measure. Alpha reliabilities are included in Table 1.

4.2.2. Means and correlations

Once we had determined which items belonged on the scale, and had verified their reliability, we analyzed the answers people had given to those items. Figure 1 shows the means and standard deviations for each factor. Table 2 shows the intercorrelations among the eight factors. Item responses are on a seven-point scale, making 4 the midpoint, labeled "neither agree nor disagree." Only mean scores for the Proximal-D factor were not significantly different from the midpoint ($M = 4.06$, $t(257) = .75$, *ns*). Mean scores for Fatalism and Distal-D were significantly below midpoint (Fatalism: $M = 2.95$, $t(257) = -17.10$, $p < .001$; Distal-D: $M = 3.08$, $t(257) = -10.89$, $p < .001$), demonstrating *dis*agreement. Mean agreement for all four formulations of free will and for Moral Responsibility (MR) was significantly above midpoint (ATDO-Incompatibilism: $M = 4.27$; ATDO-Compatibilism: $M = 5.06$; SH-Incompatibilism: $M = 4.86$; SH-Compatibilism: $M = 5.12$; MR: $M = 5.73$). This indicates that, on average, participants tended to agree with *all* these views. Moreover, except for the MR scale, these scales had the logic of possible worlds built into them,[18] so the results indicate that participants tended to be both natural compatibilists and incompatibilists about *both* notions of free will (ATDO and SH).

4.2.3. Correlations for free will and moral responsibility sub-scales

We found distinct factors for compatibilist and incompatibilist versions of both ATDO and SH views of free will. For ATDO, there was a significant difference in mean agreement between the compatibilist and incompatibilist factors, showing that agreement with the compatibilist version is significantly greater ($M_{diff} = .79$, $t(255) = 8.75$, $p < .001$). The same pattern also holds for SH free will ($M_{diff} = .26$, $t(254) = 4.53$, $p < .001$), although the magnitude of the difference is much smaller.

Although patterns in mean agreement were the same for ATDO and SH, the correlation patterns were different. Positive correlations show that the more participants agree with items on one factor, the more they will agree with items on

Table 1. Final item set EFA results. Exploratory Factor Analysis with Maximum Likelihood estimation and Varimax rotation.

	Factor							
Item	Proximal Determinism (α = .88)	Distal Determinism (α = .94)	ATDO Incompatibilism (α = .84)	ATDO Compatibilism (α = .82)	SH Incompatibilism (α = .72)	SH Compatibilism (α = .86)	Moral Responsibility (α = .88)	Fatalism (α = .79)
1_01	**.696**	.286	−.166	.065	.099	−.047	−.011	.050
7_01	**.757**	.268	−.078	.113	.034	−.099	.083	.089
2_07	**.572**	.292	−.157	.132	−.114	−.015	−.079	.219
5_07	**.795**	.180	−.163	.121	.009	.056	−.073	.184
9_06	**.638**	.203	−.222	.269	.053	.038	.042	.157
2_02	.265	**.788**	−.044	−.091	−.096	−.086	−.200	.199
1_07	.301	**.707**	.003	−.003	−.108	−.073	−.219	.169
6_03	.187	**.799**	−.094	−.013	−.104	−.022	−.180	.228
7_06	.250	**.806**	−.041	−.097	−.087	−.092	−.186	.201
10_07	.306	**.760**	−.087	−.116	.032	−.128	−.162	.240
2_08	−.181	.001	**.763**	−.006	.081	−.005	.046	.027
4_05	.023	.060	**.745**	−.150	.060	−.014	−.022	.023
9_07	−.211	−.088	**.686**	−.040	.008	.056	.159	−.005
8_03	−.205	−.125	**.665**	−.030	.089	.141	.036	.143
9_03	−.044	−.043	**.678**	−.107	−.064	.164	−.004	.109
5_05	−.009	−.070	−.137	**.704**	.223	.071	.145	−.184
7_07	.074	−.135	−.017	**.575**	.197	.146	.202	−.025
10_05	.137	−.104	−.153	**.692**	.078	.094	.131	−.047
6_02	.182	.051	−.034	**.595**	.117	−.050	.089	−.194
1_04	.120	−.019	−.059	**.739**	−.049	.125	.058	−.069
4_08	.009	−.077	.065	.048	**.559**	.184	.081	.009
6_04	.061	−.082	.052	.287	**.437**	.209	.045	−.151
10_03	.039	−.019	.074	.136	**.668**	.088	.059	−.134
8_01	−.010	−.020	−.055	.076	**.537**	.184	.225	.078
2_03	.008	−.165	.179	.181	.360	**.592**	.240	−.002
4_03	.023	−.079	.127	.050	.289	**.612**	.389	−.172

(*Continued*)

Table 1 – *Continued*

	Factor							
Item	Proximal Determinism (α = .88)	Distal Determinism (α = .94)	ATDO Incompatibilism (α = .84)	ATDO Compatibilism (α = .82)	SH Incompatibilism (α = .72)	SH Compatibilism (α = .86)	Moral Responsibility (α = .88)	Fatalism (α = .79)
5_03	−.106	−.049	.058	.164	.275	**.675**	.271	−.036
7_02	−.005	−.112	.136	.113	.199	**.727**	.232	−.041
2_03	.008	−.165	.179	.181	.360	**.592**	.240	−.002
6_05	.019	−.150	.069	.111	.024	.240	**.858**	−.175
1_09	.077	−.143	−.011	.089	.081	.141	**.797**	−.041
4_04	−.075	−.182	.116	.119	.183	.061	**.699**	−.257
2_04	−.023	−.166	.031	.165	.099	.190	**.700**	−.095
7_03	−.084	−.183	.054	.213	.175	.215	**.599**	−.045
1_05	.224	.287	.158	−.128	.042	−.011	−.095	**.672**
9_04	.098	.373	.047	−.157	.055	−.101	−.038	**.539**
5_04	.150	.111	.117	−.089	−.199	.019	−.160	**.532**
10_06	.190	.365	.044	−.145	.066	−.259	−.114	**.526**
7_04	.090	.204	.048	−.154	−.114	−.008	−.213	**.505**

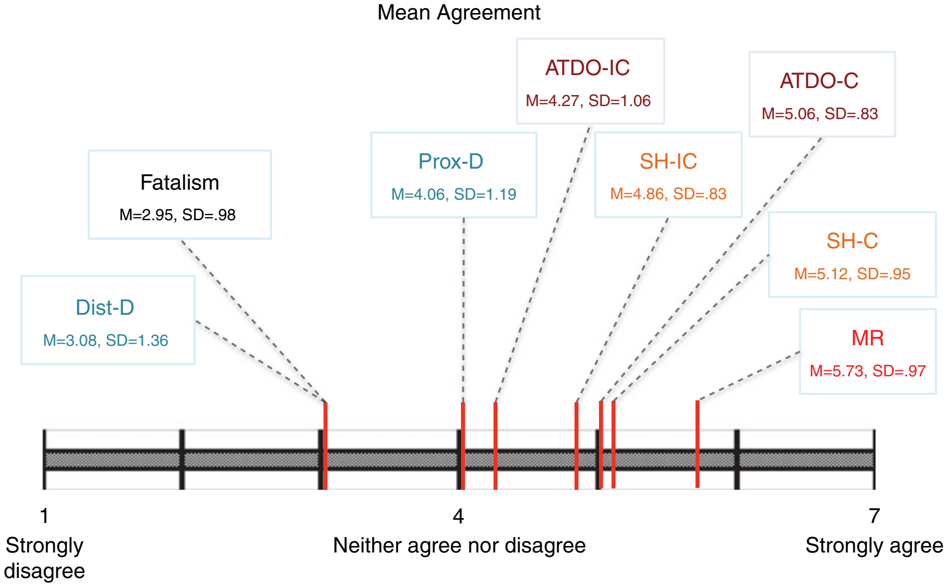

Figure 1. Mean agreement.

Table 2. Matrix of correlations among the eight sub-scales.

	Proximal-D	Distal-D	ATDO-IC	ATDO-C	SH-IC	SH-C	MR	Fatalism
Proximal-D	–	.54***	−.29***	.21***	.07	−.09	−.06	.40***
Distal-D		–	−.15**	−.16***	−.13*	−.33***	−.42**	.60***
ATDO-IC			–	−.16**	.05	.23***	.12	.08
ATDO-C				–	.33***	.31***	.33***	−.25***
SH-IC					–	.46***	.31***	−.09
SH-C						–	.52***	−.27***
MR							–	−.32***
Fatalism								–

another factor. Negative correlations show that agreement to one position brings disagreement with the other. For ATDO, compatibilism was negatively correlated with incompatibilism ($r = -.16$, $p < .01$). For SH, by contrast, the compatibilist and incompatibilist factors were strongly positively correlated ($r = .46$, $p < .001$). ATDO-Incompatibilism was negatively correlated with both Proximal-D and Distal-D ($r = -.29$, $p < .001$ and $r = -.15$, $p < .01$, respectively). ATDO-Compatibilism was positively correlated with Proximal-D ($r = .21$, $p < .001$), but negatively correlated with Distal-D ($r = -.16$, $p < .05$). Neither form of sourcehood free will was correlated with Proximal-D, but both were negatively correlated with Distal-D (SH-Compatibilism: $r = -.33$, $p < .001$, SH-Incompatibilism: $r = -.13$, $p < .05$).

All four forms of free-will intuition were positively correlated with Moral Responsibility (MR), although these correlations were higher for SH views than for ATDO views on both compatibilist and incompatibilist variants. MR was more strongly correlated with ATDO-Compatibilism ($r = .33$, $p < .001$) than with ATDO-Incompatibilism ($r = .12$, $p < .05$). The same pattern held for SH-Compatibilism and SH-Incompatibilism ($r = .52$, $p < .001$ and $r = .31$, $p < .001$, respectively). MR and Proximal-D were uncorrelated ($r = -.06$, $p = .31$), but MR was strongly negatively correlated with Distal-D ($r = -.42$, $p < .001$).

4.2.4. Correlations for determinism and fatalism sub-scales

Proximal-D and Distal-D were highly correlated ($r = .54$, $p < .001$), and both were highly correlated with Fatalism, though the relationship was stronger for Distal-D than for Proximal-D ($r = .60$, $p < .001$ versus $r = .40$, $p < .001$). Like Distal-D, Fatalism was strongly negatively correlated with MR ($r = -.32$, $p < .001$).

5. General Discussion

5.1. Formulating the Scale: What the Factor Analysis Shows

The factor analysis tells us which items belong on the scale in the first place, and it does so by telling us whether participants distinguish various kinds of statements—for example, incompatibilist and compatibilist statements expressing sourcehood notions of free will. The strength of the scale methodology is that it succeeds in distinguishing such views even when participants agree with both of them, and even when they do so to the same degree. For instance, participants in our surveys clearly distinguished compatibilist and incompatibilist sourcehood statements, but they also agreed, and to nearly the same degree, with statements expressing each view. Regarding the ability to do otherwise, participants also distinguished compatibilism from incompatibilism. Yet while they again tended to agree with statements expressing each view, this time they didn't do so to the same degree. Rather, we found that when free will was framed as the ability to do otherwise, participants strongly favored compatibilism.

The fact that participants distinguished compatibilist from incompatibilist statements, but agreed with both despite their logical incompatibility, is exactly the kind of information that the Conflict Method misses, and that our method is suited to accessing. Moreover, because all these statements expressed coherent philosophical positions, the best explanation for the emergence of any particular factor is that participants' shared with each other a single conception of that philosophical position. This makes participants' responses to the items in our surveys relevant to philosophical issues in a way that responses to previous free-will scales were not.

5.2. Explaining Away Incompatibilist Intuitions

Establishing which items belong on the FWIS, and what the factors are, enables us to speak directly to recent controversies in experimental philosophy about how notions like determinism should be operationalized. Indeed, there is a sense in which the development phase for scales is more useful in experimental philosophy than it is in psychology. A lot of data must be collected in developing a scale, and for most psychological purposes this data is of no further interest once the scale has been designed and validated. In experimental philosophy, by contrast, data collected in the design phase serve another purpose, since they tell us how to present people with questions they will understand in philosophically relevant ways.

For example, it is an empirical question whether participants misinterpret—in some way or another—statements of determinism. As Nahmias and Murray (2011) note, if respondents do misinterpret determinism, then their responses won't be about the compatibility of *determinism* and free will, so they will be irrelevant to the question whether people are natural compatibilists. Nahmias and Murray defend natural compatibilism, and they exploit this potential for misinterpretation as a way of explaining away the incompatibilist intuitions identified by philosophers such as Nichols and Knobe (2007). Nahmias and Murray argue that a significant number of what *seem* to be incompatibilist intuitions actually are not. Incompatibilist intuitions often result from misunderstanding determinism in one of two ways, each of which is a form of *bypassing*.

One form of bypassing is mechanistic. According to Nahmias and his collaborators (e.g., 2005), when statements of determinism are framed neurologically rather than psychologically. participants have a greater tendency to judge agents as not free or responsible. This misinterpretation of determinism occurs when participants are told that an agent's decision was "completely caused"[19] by his or her neural processes, but not when it was completely caused by his or her thoughts and desires.[20] The suggestion is that when an agent's decision is described neurally, the role of the agent gets bypassed. This doesn't occur, they claim, when the decision is described psychologically. Clearly, however, determinism doesn't imply bypassing of deliberation, since determinism applies equally to neurological and psychological causes of decisions. As a result, Nahmias and Murray suggest, participants' intuitions are often not about the incompatibility of determinism and free will, but about the incompatibility of *mechanistic descriptions of decisions* and free will.

We think the mechanistic bypassing hypothesis is plausible. Nonetheless, our data show that the conflation posited in this form of bypassing simply doesn't occur: outside the context of compatibility questions, statements of determinism framed neurally are just as likely to be interpreted by participants *as statements of determinism* as are statements framed psychologically, as we now explain.

In collecting our data, we included determinism items phrased both neurally and psychologically. A typical neural item read as follows:

> At 2:07 p.m. on Monday, Lucy decided to switch on the TV; the *specific chemical reactions and neural processes occurring in Lucy's brain* [emphasis added] at the time of her decision ... made it the case that her decision had to happen the way it did.

By contrast, a typical psychological item read as follows:

> Last Friday, Olivia decided to switch on the TV; Olivia's *thoughts and desires* [emphasis added] at the time of her decision ... completely caused that decision.

If participants were more likely to interpret psychological items as statements of determinism, this would predict that loadings for these items on our Proximal-D factor would be significantly higher than loadings for neural items. Moreover, if the difference between the two kinds of statement were completely clear to participants, then neural and psychological wordings would form distinct factors in our analysis.

This wasn't what we found. Neural and psychological wordings both loaded highly on the same factor—Proximal-D—and these loadings were no higher for items that referred to thoughts and desires than they were for items referring to neural causes (mean for psychological items: .668; for neural items: .727). Thus, each sort of item contributed equally to the emergence of the Proximal-D factor, due to the observed similarity among responses across all items of this kind. Given that the results obtained by Nahmias and colleagues don't appear to be due to a general tendency to misinterpret determinism as mechanism, we think they must be due, instead, to the way in which people resolve conflicts of intuition when placed in a forced-choice experiment.

The second form of bypassing is said to occur when participants misinterpret determinism as fatalism, the thesis that some events will happen no matter what. Fatalism implies that certain actions will happen regardless of whatever deliberations lead up to them: our behavior isn't caused by our deliberations, so our agentive capacities are bypassed. Nahmias and Murray (2011) predicted that the greater the extent to which participants confused determinism with fatalism, the less they would tend to judge that agents are free or morally responsible, and their findings support this prediction.

As noted, however, Nahmias and Murray employed the Conflict Method in the experiments they conducted, placing participants in the context of a philosophical debate before eliciting responses.[21] By contrast, our method investigated participants' interpretation of the relevant positions *prior* to their consideration of compatibility questions. What we found was that Nahmias and Murray's results may have depended on the way in which determinism was presented in their experiments. In our studies, participants did tend to conflate determinism with fatalism, just as Nahmias and Murray suggest, but they did so only to a certain degree, and only in a certain respect.

First, statements of fatalism formed their own factor, which was distinct from both the Proximal-D and the Distal-D factors. This shows that there were at least some systematic differences in participants' understanding of all three types of statement. That said, the Distal-D factor was highly correlated with the Fatalism factor ($r = .60$, $p < .001$). When determinism was described in terms of distal causes of action, such as the Big Bang, responses to statements of fatalism were indeed similar to responses regarding determinism. Further, participants agreed with Distal-D and Fatalism items to almost exactly the same degree. Across more than two hundred individuals, the average level of (dis)agreement with Fatalism items was 2.95, while for Distal-D items it was 3.08 (Figure 1). This striking similarity, along with the strong positive correlation between them, supports the claim that fatalism indeed tends to be conflated with distal-determinism. So, while these positions are distinguished to some degree in the minds of the folk, the distinction is hazy.

However, distinct factors also emerged for Fatalism items and statements of determinism citing *proximal* causes of action—for example, "people's thoughts and desires just prior to their decisions"—and here the distinction was much less hazy. The correlation between these factors was substantially lower ($r = .40$, $p < .001$).[22] Moreover, as Figure 1 shows, while respondents clearly disagreed with Fatalism statements (on average), they neither agreed nor disagreed with Proximal-D statements. Many of the Conflict Method studies carried out so far describe determinism primarily in terms of distal causes. This leaves open the possibility that even if the folk tend to conflate Fatalism with determinism described in terms of distal causes, they may not do so when determinism is described in terms of proximal causes. That is, there may still be respects in which the folk are incompatibilists about free will and *determinism*, properly understood. Alternatively, studies that describe determinism in terms of proximal causes may end up yielding more reliable compatibilist results, by avoiding the tendency to conflate Distal-D and Fatalism. Finally, as we explain below, there are some compatibility questions for which it is Proximal-D, not Distal-D, that raises its head as a *threat* to free will.

In all this, we learned two important lessons for future studies. First, it doesn't matter whether determinism is described neurally or psychologically—at least when asking for agreement to scale items.[23] Second, when designing experiments that ask the folk to assume determinism for the sake of making a philosophical judgment, it makes a big difference whether determinism is described proximally or distally. As we explain below, this complicates questions about how intuitions about determinism bear on philosophical debates. Though Proximal-D and Distal-D may be psychologically distinct in the minds of the folk, they are not logically distinct with regard to philosophical questions about compatibility. It is determinism itself that is thought to be logically incompatible with freedom and responsibility, and it makes no philosophical difference whether the events that determine an action occurred recently, or a long time ago.[24]

Another thing we discovered is that folk notions of fatalism are wildly heterogeneous. Some of our fatalism items were fashioned to express logical variations of this position that are of interest to philosophers. Yet we also included items expressing more colloquial usages of 'fatalism'; for example, "if it was in the stars that Robert was going to kill his father, then Robert was going to kill his father no matter what." This notion of fatalism is distinct from more rigorous philosophical statements like, "Susan's decision to have juice had to happen the way it did, no matter what thoughts and desires were going through her mind prior to her decision" (which expresses the idea that no matter what the past or the laws might have been, Susan's action had to occur). We included the colloquial items because everyday uses of 'fatalism' seem to express the idea that there is an underlying teleology to the way in which some events unfold, and we wondered whether there was anything more to the folk notion of fatalism than what some psychologists call "promiscuous teleology" (Kelemen & Rosset, 2009).

In the end, participants didn't register these differences. Average loadings on the Fatalism factor were slightly higher for the more rigorous statements ($M = .579$) than for the colloquial wordings ($M = .519$), but this difference isn't important. Loadings for some colloquial items were easily high enough for them to be included on the final version of the Fatalism sub-scale. We conclude that the folk notion of fatalism may be vague and heterogeneous, which could partially explain the strong correlation between Fatalism and Distal-D on our final scale ($r = .60$, $p < .001$). That is, since the folk notion of fatalism appears so vague in the first place, it might be easy for people to confuse it with certain wordings of determinism.

5.3. Putting the Scale to Use: What the Means and Correlations Show

Whereas the factor analysis is done in the course of formulating the scale, calculating the means and correlations is a way of *using* the scale. Once we identified the factors, the next step was to use the loadings for each item to select the best four or five items for the purposes of expressing the philosophical position represented by that factor. These items were then declared the "official" items comprising the sub-scale for that factor. Once we had identified which items would comprise the final scale, we were in a position to use the responses we had already collected for those items. This allowed us to examine mean levels of agreement to the positions expressed by each sub-scale, along with the correlations between levels of agreement for the positions represented on the scale. We have already mentioned some of these findings, but the means and correlations shown in Figure 1 and Table 2 also tell us much more about natural compatibility.

Most importantly, we found that participants in our studies tended to be *both* compatibilists *and* incompatibilists. For instance, participants tended to agree with statements expressing a compatibilist conception of sourcehood free will. Yet they also agreed with statements expressing an *in*compatibilist conception of sourcehood freedom, and to a similar extent (Figure 1). Thus, all else being equal, when respondents weren't placed in an experimental situation asking them to resolve a psychological conflict between two logically conflicting intuitions, they possessed both incompatibilist and compatibilist intuitions regarding sourcehood, and they didn't exhibit a strong preference for either one.[25] Our data thus show that even when we focus specifically on sourcehood notions of free will, it makes no sense to claim that the folk are exclusively either compatibilists or incompatibilists. This state of affairs couldn't have been identified simply by running experiments that ask individuals to *choose between* compatibilism and incompatibilism.[26]

Regarding ability-to-do-otherwise (ATDO) notions of free will, participants also agreed with both compatibilist and incompatibilist statements. However, here we found a clear preference for compatibilism (Figure 1). We take this to predict that, when faced with a psychological conflict between compatibilist and incompatibilist ATDO statements, most participants will resolve in favor of compatibilism. In an important sense, then, folk intuitions *are* compatibilist here, even though participants *also* endorsed ATDO-Incompatibilism statements. Again, the task for future research is to administer the scale along with experiments, to identify the specific contexts in which people favor compatibilism or incompatibilism.

We also discovered an interesting clustering of intuitions around two different reasons people have for being concerned about the compatibility of determinism and free will, a clustering that is reflected in the philosophical literature as well. Almost all parties to free-will debates once agreed that agents must be able to do otherwise to be responsible. In recent decades, many compatibilists (in particular) have abandoned this idea. Sourcehood theorists claim instead that responsibility requires being—in a relevant sense—the *source* of one's actions. As Table 2 shows, endorsement of sourcehood (SH) items was highly correlated with endorsement of Moral Responsibility (MR) items, especially for compatibilist sourcehood (MR and SH-Compatibilism: $r = .52$, $p < .001$; MR and SH-Incompatibilism: $r = .31$, $p < .001$). When free will is conceived as the ability to do otherwise, however, the corresponding relationship is much weaker. While we found a significant correlation between MR and ATDO-Compatibilism ($r = .33$, $p < .001$), this correlation was substantially lower than that between MR and SH-Compatibilism. More importantly, we found no correlation at all between ATDO-Incompatibilism and MR ($r = .12$, $p = .62$). To the extent that there is any meaningful relationship between intuitions about responsibility and the ability to do otherwise, it appears to be due to compatibilist intuitions. Even so, compatibilist intuitions about responsibility are more strongly related to *sourcehood* intuitions than to intuitions about the ability to do otherwise.

This suggests a clear divide among folk intuitions between *moral* and *modal* reasons for interpreting determinism as a threat to free will. The modal cluster of intuitions is represented on the FWIS by the two ATDO sub-scales, while the moral cluster is captured by the two SH sub-scales, together with the MR sub-scale. The modal cluster doesn't interact as strongly with responsibility as traditional assumptions might suggest. This indicates that the move among compatibilist philosophers away from the ability to do otherwise as a condition on responsibility, and instead toward sourcehood conditions, is a move in the direction of folk intuitions.

Even more strikingly, we found an interesting interaction between the moral and modal clusters and the two forms of determinism. When proximal causes of action were cited (e.g., "people's thoughts and desires just prior to their decisions"), interesting relationships emerged between determinism and the modal cluster. Yet when distal causes were cited (e.g., the Big Bang), the interesting relationships were between determinism and the moral cluster. More specifically, we found a negative correlation between Proximal-D and ATDO-Incompatibilism ($r = -.29$, $p < .001$), and a positive correlation between Proximal-D and ATDO-Compatibilism ($r = .21$, $p < .005$). However, we found no significant correlation between Proximal-D and either SH (whether Compatibilist or Incompatibilist) or MR (MR: $r = -.06$, SH-Compatibilism: $r = -.09$; SH-Incompatibilism: $r = .07$). Thus, Proximal-D appears to affect modal intuitions regarding the ability to do otherwise, but it has no corresponding effect on the moral cluster of intuitions.

By contrast, Distal-D affected the moral cluster more than the modal cluster. Distal-D was negatively correlated with both ATDO factors, but these correlations were extremely weak (ATDO-Compatibilism: $r = -.16$, $p < .05$; ATDO-Incompatibilism: $r = -.151$, $p < .05$). Moreover, the relationship between Distal-D and SH-Incompatibilism was also weak ($r = -.13$, $p < .05$). By contrast, the relationships between Distal-D and the other two factors in the moral cluster were much stronger.

First, Distal-D was strongly negatively correlated with MR ($r = -.42, p < .001$). Recall that, within the moral cluster, MR is strongly correlated with SH-Compatibilism ($r = .52$, $p < .001$). Accordingly, it isn't surprising that there is a strong negative correlation between Distal-D and SH-Compatibilism ($r = -.33$, $p < .001$) as well. Indeed, while SH-Compatibilism is more highly *positively* correlated with MR than it is with any other factor, it is also more highly *negatively* correlated with Distal-D than with any other factor. Compatibilist intuitions about sourcehood appear to be intimately connected to those about responsibility, but they are also psychologically incompatible with Distal-D.[27] This suggests that what is distinctively compatibilist about sourcehood intuitions is the compatibility of sourcehood free will and *Proximal-* D. This seems to be because (natural) sourcehood compatibilists care more about responsibility than anything else, and retain a natural *in*compatibilist tendency about the relationship between Distal-D and MR.

Thus, while Proximal-D is more often seen as a threat to modal concerns regarding the ability to do otherwise, Distal-D is more often seen as a threat to moral concerns regarding responsibility.

5.4. Explaining Away Compatibilist Intuitions

Just as Nahmias and his collaborators provide an error theory for incompatibilist intuitions, Nichols and Knobe (2007) suggest an error theory for *compatibilist* intuitions. According to their view, the folk naturally have an incompatibilist theory of responsibility, but their application of this theory is sometimes clouded by affective responses. In "abstract" conditions, which elicit theoretical cognition, people tend to be incompatibilists about responsibility, while in "concrete" conditions, which elicit an affective response, people tend to be compatibilists. Nichols and Knobe claim that the affective response induces a performance error, by interfering with participants' ability to reason correctly in applying their theory of responsibility.

Our results don't speak directly to Nichols and Knobe's error theory, partly because we didn't test for effects of affect or concreteness. For all the FWIS tells us, Nichols and Knobe's error theory might be correct, in that it identifies *one* way in which conflicts of intuition about responsibility get *resolved* by the folk. What the FWIS does tell us—and Nichols and Knobe's experiments don't—is what folk intuitions about responsibility look like *prior* to participants' being put in forced-choice experimental situations. We found that the conception of free will most intimately related to intuitions about responsibility was a compatibilist notion, not an incompatibilist one. We think this suggests that Nichols and Knobe's findings are more limited than they might otherwise appear. Let us explain.

Nichols and Knobe's results support the view that the folk are natural incompatibilists about responsibility. According to our findings, however, this compatibility question is psychologically distinct from other compatibility questions about free will. Even if Nichols and Knobe's data show that people are natural incompatibilists about responsibility, our data show that people are natural compatibilists in other senses. Most strikingly, incompatibilist intuitions about *responsibility*[28] co-exist in the minds of the folk alongside compatibilist intuitions about *free will—even for the notion of free will most immediately related to intuitions about responsibility.*

Second, our results are consistent with Nichols and Knobe's incompatibilist results, in that we too find that MR responses are negatively correlated with Distal-D responses ($r = -.42$, $p < .001$). Further, we found that the greater the extent to which people agreed with MR statements, the more they also agreed with SH-Incompatibilism ($r = .31$, $p < .001$). Thus, since sourcehood free will is part of the moral cluster, rather than the modal cluster, one might think that these distinct forms of incompatibilism are psychologically related. That is, perhaps people are natural incompatibilists about both responsibility and sourcehood free will because they take sourcehood to be what is required for responsibility.

However, we found that the correlation between MR and SH-*Compatibilism* is even higher than that between MR and SH-Incompatibilism, and by a considerable margin (.31 versus .52). This predicts that, across a range of studies using the Conflict Method, substantially more people will resolve conflicts between compatibilist and incompatibilist sourcehood intuitions in favor of compatibilism. So, while our results do identify a certain variety of incompatibilist intuition about responsibility, it is of a constrained and limited sort. Even if someone has (partly) incompatibilist intuitions about MR, this doesn't mean she will also have incompatibilist intuitions about free will.[29] If this is the extent of the incompatibilist theory that Nichols and Knobe can attribute to the folk, then it is much weaker than the one that they apparently have in mind.

This suggests a further sense in which Nichols and Knobe's results are limited: they seem to apply only to Distal-D. Our results suggest that individuals who resolved the conflict in Nichols and Knobe's experiments in favor of incompatibilism might have chosen compatibilism instead, had the vignettes cited proximal rather than distal causes. Of course, we have no evidence indicating that incompatibilists about MR and Distal-D *would* respond as compatibilists about MR and Proximal-D. However, the kind of natural incompatibilism that Nichols and Knobe identify is based only on intuitions from the moral cluster, and our data show that this kind of natural incompatibilism doesn't automatically come along with incompatibilist intuitions from the modal cluster, where proximal causes are of primary importance. Moreover, Proximal-D and Distal-D are not distinct philosophical positions, even if they function differently in the minds of the folk. Thus, before we can conclude that the folk are natural incompatibilists about MR and determinism, it remains to be shown that they are natural incompatibilists about MR and *Proximal*-D as well. If not, then it is unclear what the *philosophical* content of the incompatibilist intuitions identified by Nichols and Knobe is supposed to be.

Just as the data Nahmias and Murray present in support of their error theory for incompatibilist intuitions may be limited to Distal-D, the support that Nichols and Knobe provide for *their* error theory is also limited in scope. What this reveals is not that there is anything wrong with Nichols and Knobe's studies, but rather that the kind of natural incompatibilism demonstrated in their studies is, at best, of a very specific form. These limitations leave open the possibility that folk intuitions are both compatibilist and incompatibilist at the same time.

6. Concluding Remarks

The Conflict Method has produced results supporting both sides of the natural compatibilism debate, leading experimental philosophers on each side to propose error theories explaining away opponents' results. Our findings suggest that these error theories have been proposed too soon. Even if folk intuitions are compatibilist or incompatibilist in a particular experiment, this is no reason to think that folk intuitions are *consistently* compatibilist or incompatibilist (compare Doris, Knobe, & Woolfolk, 2007). Experiments simply aren't the right tool for measuring judgments across a wide range of circumstances and contexts, and there is no a priori reason to assume that folk intuitions will be consistent across all these contexts.

While experiments employing the Conflict Method are capable of telling us how people resolve conflicts of intuition in specific contexts, scale methodology allows us to "zoom out," and to gain an overview of the ways in which responses *across* experimental contexts are related. This overview shows us that empirical questions about natural compatibility are much more complicated than has been recognized. So before we assume that certain intuitions need to be explained away, we should seriously consider the empirical possibility that folk intuitions aren't especially coherent or logically consistent. This possibility can't be explored if coherence and logical consistency are already built into the design of the questions that we present to the folk.

Our primary aim here has been to convince others—philosophers and psychologists alike—that uncovering what folk intuitions are like *prior* to participants' being put in forced-choice experiments can make experimental results more powerful, by illustrating what the *basic* intuitions are that participants bring to these choices, thereby showing how data across many different experiments are related.

Appendix A

Item numbers correspond to the numbers in Table 1 and are presented in the same order.

Proximal Determinism

1_01 At 2:07 p.m. on Monday, Lucy decided to switch on the TV; the specific chemical reactions and neural processes occurring in Lucy's brain at the time of her decision—in accord with the laws of nature—made it the case that her decision had to happen the way it did.

7_01 At 8 p.m. yesterday evening, Emma decided to watch a movie; the specific chemical reactions and neural processes occurring in Emma's brain at the time of her decision—in accord with the laws of nature—completely caused that decision.

2_07 If a week from now I decide to eat pasta for dinner, my thoughts and desires just prior to my decision—in accord with the laws of nature—will make it the case that I have to decide to have pasta for dinner.

5_07 On Saturday evening, Michael decided to watch a movie; Michael's thoughts and desires at the time of his decision—in accord with the laws of physics—made it the case that his decision had to happen the way it did.

9_06 People's thoughts and desires just prior to their decisions—in accord with the laws of nature—make it the case that their decisions have to happen the way they do.

Distal Determinism

2_02 At exactly two o'clock last Tuesday, Jack decided to eat a pizza; the state of the universe millions of years ago and all the subsequent events right up until Jack's decision—in accord with the laws of nature—made it the case that his decision had to happen the way it did.

1_07 The state of the universe millions of years ago and all the events ever since then—in accord with the laws of nature—make it the case that my decisions have to happen the way they do.

6_03 Imagine that at 7:55 p.m. yesterday evening, you decided to watch a movie; the state of the universe millions of years ago and all the events after that right up until your decision—in accord with the laws of nature—made it the case that your decision had to happen the way it did.

7_06 If a week from now Mike decides to buy a new guitar, the state of the universe millions of years ago and all the subsequent events right up until Mike's decision—in accord with the laws of nature—make it the case that he will have to decide to buy a new guitar.

10_07 If a week from now I decide to have a soda with lunch, the state of the universe millions of years ago and all the subsequent events right up until my decision—in accord with the laws of nature—make it the case that I will have to decide to have a soda with lunch.

Ability-To-Do-Otherwise Incompatibilism

2_08 I could have decided to buy a different detergent than I actually bought, but I would have decided to do so even if none of my desires or thoughts at the time had been different.

4_05 Amelia bought the *Washington Post* because she prefers it to the *New York Times*; she could have bought the *New York Times* instead and she might have done so even if her preferences and the situation had been exactly the same.

9_07 I could have bought a different dish soap than I actually bought, and I might have done so even if none of my preferences or desires had been different.

8_03 I chose the noodle dish, but I could have chosen the rice dish instead, even if everything at the moment of my choice—including my thoughts and desires—had been exactly the same.

9_03 Emily could have taken the job in London instead of the job in San Francisco, and she might have decided to do so even if none of her desires or thoughts had been different.

Ability-To-Do-Otherwise Compatibilism

5_05 I might have taken the job in Chicago instead of the job in Atlanta, but I would have done so only if my thoughts or desires had been different as I made the decision.

7_07 Audrey might have chosen to take the job in St. Louis instead of the job in Toronto, but only if she had wanted the job in St. Louis more.

10_05 I might have decided to take the job in New York instead of the job in San Francisco, but only if something at the time of my decision had been different—for instance, only if I'd had different desires, or different considerations had come to mind.

6_02 I could have decided to buy a different detergent than I actually bought, but I would have decided to do so only if my thoughts or desires at the time had been different.

1_04 Henry might have decided to take the job in San Francisco instead of the job in London, but he would have done so only if something at the time of his decision had been different.

Sourcehood Incompatibilism

4_08 Even when I decide what to do on the basis of my own values, that's not enough for me to be the ultimate source of my decisions; I must also have had the final say about what my values were in the first place.

6_04 Even when my moral reasoning is able to affect the desires I act on, that's not enough for my decisions to be up to me; it must also be true that at some point in the past I had the final say about how I responded to moral reasoning and whether I accepted such reasoning as my own.

10_03 Even when Owen decides what to do on the basis of his own reasons, that's not enough for him to be the ultimate source of his actions; he must also have had the final say about how he responds to such reasons.

8_01 Even when Sophia does what she wants and identifies with it, that's not enough for her to be the ultimate source of her actions; it must also be true that she had the final say about what she wanted in the first place.

Sourcehood Compatibilism

2_03 As long as James does what he wants and identifies with it, that's enough by itself for him to be the ultimate source of his actions; in other words, that's enough for his actions to be "up to him."

4_03 As long as Hannah decides what to do on the basis of her own reasons, that's sufficient for her to be the ultimate source of her actions; in other words, that's enough for her actions to be "up to her."

5_03 As long as my moral reasoning can affect the desires I act on, that's sufficient for my being the ultimate source of my actions; in other words, that's enough for my actions to be "up to me."

7_02 As long as I decide what to do on the basis of my own values, that's enough by itself for me to be the ultimate source of my decisions; in other words, that's enough for my actions to be "up to me."

Moral Responsibility

6_05 Matt dived into the swimming pool and rescued a drowning child, so he deserved the praise he later received for his action.

1_09 When a person deliberately helps another person that she doesn't really have to help, she deserves whatever reward she later receives for this action.

4_04 Freya stole the car just for fun, and she knew that stealing it was wrong, so she deserves the blame she later received for this action.

2_04 Even though she was under no obligation to donate money, Julia donated 20% of her income to the victims of the earthquake, so she deserved the praise she later received for her action.

7_03 When someone deliberately harms another person, she deserves whatever punishment she later receives for her action.

Fatalism

1_05 Susan's decision to have juice had to happen the way it did, no matter what thoughts and desires were going through her mind prior to her decision.

9_04 Anna's decision to make a cup of tea had to happen the way it did, regardless of the thoughts and desires she had at the time.

5_04 Sartre always knew he'd become famous, so he was going to become famous no matter what.

10_06 Josh's decision to eat steak for dinner had to happen the way it did, no matter what chemical and neural processes were going on in his brain at the time.

7_04 On the night Rachel was born, her great-grandfather had a dream in which Rachel was surrounded by her own seven grandchildren, so it was inevitable that Rachel would grow up to have seven grandchildren of her own.

Acknowledgements

We thank Joshua Knobe, Shaun Nichols, Derk Pereboom, Fiery Cushman, and the participants at the Experimental Philosophy of Free Will Workshop and Bootcamp at Yale University, August 12–13, 2011, for their helpful comments. Thanks also to Matt Bedke, Joe Campbell, Eddy Nahmias, and three anonymous reviewers at this journal.

Notes

[1] We discuss these error theories in section 5.2 and section 5.4.

[2] In case it appears we are entering the same debate we suggest is misguided, consider the difference between the general project of understanding intuitions about free will and the more specific project of showing that people are *either* compatibilists or incompatibilists. We want to understand what people's intuitions are even if they are not exclusively of either type.

[3] Even when there isn't any internal conflict (perhaps one option is clearly preferable), this is still the Conflict Method, since the method presents statements as logically conflicting.

[4] Compare this with Greene's work (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008) showing that utilitarian moral judgments are driven by controlled cognitive processes, whereas non-utilitarian moral judgments are driven by automatic emotional responses. These processes compete, and fMRI data describe the competing processes themselves. This information would be missed by a Conflict study that only recorded the outcomes of these cognitive competitions.

[5] Using a continuous variable (e.g., a Likert scale ranking) instead of a discrete *A/B* variable changes nothing, since the *degree or strength of agreement* someone feels for the "losing" basic intuition is not measured. Measuring a participant's *degree of confidence* in a choice between exclusive options differs from measuring his or her *degree of agreement* with *each* option.

[6] One might object that a "neither disagree nor agree" answer in a Conflict Method experiment *is* a way of choosing both sides. Yet such an answer is ambiguous between agreeing with neither option and agreeing with both options.

[7] For example, Nichols and Knobe (2007) seek to understand the processes that produce decision-outcomes, and they do so by designing a new set of experiments.

[8] For the same reason, Nadelhoffer, Shepard, Nahmias, Sripada, & Ross (forthcoming) have developed a philosophical free-will scale that focuses on relationships between free-will intuitions and intuitions about dualism.

[9] We say more about the fatalism items in sections 4 and 5, and about the determinism items both in those sections and later in this section.

[10] Since our ATDO items describe only proximal causes of action, they do not rule out the possibility that respondents assume indeterminism for times prior to the relevant determining causes. One might thus worry whether agreement here indicates *legitimate* compatibilist intuitions. Furthermore, if participants assume determinism only at the time of the action, and not prior to it, they would have an "improper" notion of determinism in mind, and thus wouldn't be judging ATDO as compatible with *determinism* proper. We recognize this empirical possibility, but consider it unlikely. That said, the only way to know whether such possibilities threaten our results is to pursue exactly the sort of additional empirical work we designed the scale to stimulate.

[11] Incompatibilists about the ability to do otherwise (hard determinists included) *can't* agree with this (see note 12).

[12] Incompatibilists *can* agree that agents have a compatibilist ability to do otherwise. However, incompatibilists *can't* consistently agree with our compatibilist ATDO items, since these items deny that agents can do otherwise given the actual past and laws. Likewise, compatibilists can agree (although we know of no one who has done so) that agents have an incompatibilist ability to do otherwise, yet still insist that it isn't required for anything of importance. Nevertheless, compatibilists can't consistently agree with our incompatibilist ATDO items, since these items assert what compatibilists deny: even if everything (past and laws) prior to an agent's decision were to remain fixed, the agent could have done otherwise.

[13] This tendency is largely due to Frankfurt (1969). See Timpe (2013) for an overview of the relevant literature.

[14] Standardly, there is both a control condition and an epistemic condition on moral responsibility. The sourcehood compatibilist's claim is that an action's issuing from the agent in the right way is sufficient, at least *control-wise*, for responsibility; for *all* the conditions jointly sufficient for responsibility to be met, the epistemic condition must also be satisfied.

[15] We recognize that a certain kind of compatibilist might agree with this incompatibilist claim, if she takes volitional states, not reasons-responsive states, to be important for sourcehood. Moreover, respondents' having such intuitions could explain the correlation we found between SH-Compatibilism and SH-Incompatibilism factors (section 4.2.3). In fact, we did find that the item tracking this alternative sort of SH-Compatibilism (item 2_03, appendix A) is strongly correlated with agreement to SH-Incompatibilist items citing "reasons," "reasoning," and "values" (items 4_08, 6_04, 10_03). Yet we *also* found a strong correlation between this item (2_03) and its direct incompatibilist denial (item 8_01). Thus, the hypothesis that compatibilists are *coherently* responding as incompatibilists doesn't appear to be the best explanation for the general correlation between SH-Compatibilism and SH-Incompatibilism. Nevertheless, this is precisely the sort of new empirical hypothesis that could be examined in more detail, and that the FWIS helps us to identify.

[16] Nothing important depends on how people use the term 'free will'. Our scale addresses substantive conceptual issues, not semantic issues about 'free will'. We wanted to see

whether people agree to statements that imply agreement to (in)compatibilist notions of ATDO or SH, regardless of how they are inclined to use 'free will'.

[17] A typical Proximal-D item is, "People's thoughts and desires just prior to their decisions—in accord with the laws of nature—make it the case that their decisions have to happen the way they do." A typical Distal-D item is, "The state of the universe millions of years ago and all the events ever since then—in accord with the laws of nature—make it the case that my decisions have to happen the way they do." Note that these items ask only about the actual world, so they are not directly relevant to the compatibility issue, which concerns whether we have free will in any possible world where determinism is true. By contrast, the SH and ATDO items *do* directly address this issue, since the logic of possible worlds is built into them.

[18] See note 17.

[19] This was Nahmias and colleagues' term of art for 'determined'. We tested for differences in participants' responses dependent on whether the phrase 'completely caused' was used, or instead 'had to happen' (compare Nichols & Knobe, 2007). None was found.

[20] In both these cases, of course, the decision is determined by prior states of the agent *in accord with the laws of nature*.

[21] There is a difference between placing participants in the context of a debate, and thus priming them to look for logical inconsistencies, and building consistency into the answer choices by disallowing participants to say that free will is *both* consistent *and* inconsistent with determinism (as they can on a scale). By "Conflict Method," we refer to experimental studies that do both.

[22] This correlation is still quite strong, but the important point is that it is substantially lower than the degree to which the Distal-D factor correlated with the Fatalism factor.

[23] This is consistent with its (perhaps) mattering in certain forced-choice experiments.

[24] Perhaps there *is* a philosophical difference: proximal causes may leave it open that the agent is ultimately responsible for the cause itself, while distal causes seem to preclude this (although assuming so begs the question against philosophical compatibilists). The sense in which *we* mean that there is no philosophical difference is this: each of these two formulations says that a complete description of the world (together with the laws) at some instant prior to an agent's action entails that she performs that action when she does. On this formulation, the threat to freedom or responsibility is perhaps less *salient* when determinism is describe in proximal terms, since then it specifies simply that the action is due to states of the agent, and prior causes of those states aren't mentioned. Yet that is just a psychological difference, not a philosophical one. It is also a difference our method is well-suited to uncovering.

[25] Thus, our data make no specific predictions for how respondents will, in general, resolve conflicts between compatibilist and incompatibilist sourcehood intuitions. However, see section 5.4.

[26] See note 6.

[27] By the same token, those who endorse Distal-D tend to reject MR.

[28] Unlike our ATDO and SH items, the logic of possible worlds isn't built into our MR items. Yet we do adduce some weaker, indirect evidence that folk intuitions are (partly) incompatibilist about responsibility, as we discuss immediately below.

[29] This result may come as a shock to anyone who thinks that free will *just is* whatever form of control is required for moral responsibility. We note, however, that our own findings about responsibility are also limited. Our MR sub-scale is a limited instrument, and it may take developing a new scale devoted entirely to intuitions about responsibility to get to the bottom of why participants in Nichols and Knobe's experiments resolved conflicts of intuitions between responsibility and determinism in the way they did.

References

Doris, J. M., Knobe, J., & Woolfolk, R. L. (2007). Variantism about responsibility. *Philosophical Perspectives*, *21*(1), 183–214.

Feltz, A., Cokely, E. T., & Nadelhoffer, T. (2009). Natural compatibilism versus natural incompatibilism: Back to the drawing board. *Mind & Language*, *24*(1), 1–23.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, *66*(23), 829–839.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). New York: Guilford.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*(1), 138–143.

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., & Ross, L. (forthcoming). The Free Will Inventory: A new scale on beliefs about free will, determinism, and dualism. *Consciousness & Cognition*.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*(5), 561–584.

Nahmias, E., & Murray, D. (2011). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In J. Aguilar, A. Buckareff, & K. Frankish (Eds.), *New waves in philosophy of action* (pp. 189–216). New York: Palgrave Macmillan.

Nichols, S. (2004). Folk psychology of free will: Fits and starts. *Mind & Language*, *19*(5), 473–502.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, *41*(4), 663–685.

Paulhus, D. L., & Carey, J. M. (2011). The FAD-Plus: Measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment*, *93*(1), 96–104.

Rakos, R. F., Laurene, K. R., Skala, S., & Slane, S. (2008). Belief in free will: Measurement and conceptualization innovations. *Behavior and Social Issues*, *17*, 20–39.

Stroessner, S., & Green, C. (1990). Effects of belief in free will or determinism on attitudes toward punishment and locus of control. *Journal of Social Psychology*, *130*(6), 789–799.

Timpe, K. (2013). *Free will: Sourcehood and its alternatives* (2nd ed.). New York: Bloomsbury.

Viney, W., Waldman, D., & Barchilon, J. (1982). Attitudes toward punishment in relation to beliefs in free will and determinism. *Human Relations*, *35*(11), 939–950.