Philos Stud (2012) 157:341–360
DOI 10.1007/s11098-010-9651-8

What in the world is weakness of will?

Joshua May · Richard Holton

Published online: 28 October 2010

Abstract At least since the middle of the twentieth century, philosophers have tended to identify weakness of will with *akrasia*—i.e. acting, or having a disposition to act, contrary to one's judgments about what is best for one to do. However, there has been some recent debate about whether this captures the ordinary notion of weakness of will. Richard Holton claims that it doesn't, while Alfred Mele argues that, to a certain extent, it does. As Mele recognizes, the question about an ordinary concept here is one apt for empirical investigation. We evaluate Mele's studies and report some experiments of our own in order to investigate what in the world the ordinary concept of weakness of will is. We conclude that neither Mele nor Holton (previously) was quite right and offer a tentative proposal of our own: the ordinary notion is more like a prototype or cluster concept whose application is affected by a variety of factors.

Keywords Akrasia · Weakness of will · Intention · Resolution · Experimental philosophy · Prototype concepts · Knobe effect

1 Introduction

How should we understand weakness of will? Some years ago one of the present authors published a paper arguing that the philosophical discussion had run together

J. May (✉)
Department of Philosophy, University of California, Santa Barbara, 5631 South Hall, Santa Barbara, CA 93106, USA
e-mail: jdmay@umail.ucsb.edu

R. Holton
Department of Linguistics and Philosophy, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 32-D808, Cambridge, MA 02139-4307, USA

two quite distinct notions (Holton 1999).[1] The first is the idea of *acting contrary to one's best judgement*—the idea that is to the fore in classical discussions of akrasia. The second is the idea of *over-readily revising a resolution*. It is plausible that the two typically go together: in acting contrary to their best judgments, agents will typically over-readily revise their resolutions; and conversely, in over-readily revising their resolutions they will typically act contrary to their best judgments. So it is understandable that the two ideas have been conflated. But nevertheless they are clearly distinct.

Holton could have rested content with pointing out the distinction and getting clear on the ideas involved. Rashly though, he went further. He claimed that it is just the second idea, the idea of over-ready resolution revision, that corresponds to our ordinary notion of weakness of will. The idea of acting contrary to one's best judgment he took to be a philosopher's invention, best labeled with the philosopher's proprietary term "akrasia."

He should have known better.[2] Whilst the claim enabled him to formulate his position in a neat slogan—*weakness of will is not akrasia*—it left him open to empirical refutation. For the nature of our ordinary concepts is, at least to some degree, an empirical question; and he had done no real empirical work to substantiate what he was saying. Not that it seemed that such work was needed at the time. Even 10 years ago the analysis of ordinary concepts was a leisurely affair, conducted largely by reflection, and the casual interrogation of colleagues, friends and students. Now, however, it is an altogether more rigorous business, and in a recent paper Mele (2010) has used the survey method that has become the hallmark of experimental philosophy to argue that Holton was wrong. As we understand him, Mele claims that the ordinary notion of weakness of will is disjunctive—one exhibits weakness of will *either* by acting contrary to one's evaluative judgment *or* by acting contrary to one's plan. On this view, Holton should not have been so restrictive.

Our project is to shed light on what in the world the concept of weakness of will is. Is there an ordinary notion here? If so, is it disjunctive as Mele contends? Employing some empirical methods ourselves, we argue that neither the traditional account of weakness of will, nor Holton's, nor Mele's, is quite right. Indeed, our findings suggest that no simple account phrased in terms of necessary and sufficient conditions will do the job. The ordinary notion of weakness of will is more like a prototype or cluster concept. There are core cases that possess a number of features. As these features are removed, people are less inclined to describe the resulting cases as ones of weakness of will. Akrasia and resolution-violation are indeed among these features. However, neither is sufficient on its own for an ascription of weakness of will; and other features also play a role, such as the moral valence of the action.

[1] A lightly revised version appears as ch. 4 of Holton (2009).

[2] In fact it seems that he did: "the English language is a plastic instrument" he wrote; "is it not very likely that the traditional account captures one of our uses of the expression?" But he then went on to say that he couldn't help thinking that the traditional account was straight-out wrong (pp. 257–258). He should have tried harder to resist.

2 Mele's case against Holton

In "Intention and Weakness of Will" (1999), Holton argues that the ordinary concept of weakness of will essentially involves the violation of a certain kind of intention—a *resolution*. "Resolution" is something of a technical term for Holton; it refers to an intention or plan one has to stick to a certain course of action in the face of the temptation to succumb.[3] For example, consider a smoker who sincerely admits it isn't best for her to continue smoking cigarettes but nevertheless doesn't plan to quit; she recognizes the detrimental effects of smoking but doesn't resolve to quit in spite of her evaluative judgment. Such a person is clearly being akratic. After all, "akrasia" is a term of art, and this person is by hypothesis acting contrary to what she thinks is best for her to do. But is she acting in a weak-willed manner? Is she displaying weakness of will?[4] Holton claimed not. For him, agents display weakness of will only if they violate a resolution.[5] To get a case of weakness of will then, the smoker would have to resolve to quit, but then succumb to temptation and continue. Akrasia and violations of resolutions often come and go together—we often resolve to do what we think is best for us to do—but they can come apart. Call this the *resolution account* of weakness of will.

In a recent paper, Mele (2010) argues that the ordinary concept of weakness of will involves both the notion of akratic action and the notion of intention-violation. Building on earlier work, Mele distinguishes an agent's "evaluative commitments" (roughly their judgements about which action would be best) from their "executive commitments" (roughly their resolutions). He holds that there are "traditional" or "orthodox" versions of akrasia that involve violations of one's evaluative commitments; but that there are also "nontraditional" or "unorthodox" versions of akrasia, that involve violations of one's executive commitments. After summarizing his earlier work, Mele writes:

> I did not offer full-blown analyses of akratic and enkratic action. Instead I offered sketches of conceptions of both kinds of action designed to accommodate traditional and nontraditional species of them. Are these sketches hopelessly flawed? (p. 394)

[3] It seems pretty close to the ordinary usage (consider a New Year's resolution to quit smoking). But we'll avoid making any further unsubstantiated empirical claims. Note too that the notion of temptation in play here is potentially broader than the ordinary use. We might not, for example, say that a stranded climber with his arm stuck between two rocks is *tempted* not to cut his arm off. But insofar as he resolves to cut off his arm partly to resist the urge not to, he is "tempted" in the relevant sense.

[4] Mele takes weakness of will to be a character trait. We don't want to dispute that claim here. To avoid conflict, we will follow Mele and simply say in such cases that the person "exhibits" or "displays" weakness of will (or akrasia) or is acting in a weak-willed manner (or being akratic).

[5] We say "only if" because Holton thinks violating a resolution isn't sufficient for being weak-willed—one must also do so *unreasonably*. In one formulation, Holton writes that weakness of will (or action displaying it) is "unreasonable revision of a contrary inclination defeating intention (a resolution) in response to the pressure of those very inclinations" (2009, p. 78). Since the debate between Mele and Holton doesn't revolve around this normative element, we won't focus on it in this paper and will often leave it out in characterizing Holton's view.

Mele's answer to his own question is a very definite "no." But the question that he has posed is a rather odd one, for the notions of akratic and enkratic action are philosophical terms of art. There may be an interesting question of how other theorists have used them, but this doesn't seem to be all that Mele is doing. Rather his primary focus seems to be on providing concepts that will help with the understanding of human behavior. So our tentative interpretation is that Mele thinks the account of akrasia he provides will be the most fruitful.

So far then there is no substantial issue between Mele and Holton. Disagreement comes with Mele's next claim, that our ordinary notion of weakness of will is the same as his conception of akrasia. As he puts it: "weakness of will can be displayed both in acting contrary to an evaluative commitment and in acting contrary to an executive commitment" (p. 397).

The fact that Mele only provides a "sketch" of the nature of akrasia makes it a little hard to see just what his account of the ordinary notion of weakness of will is. But his talk of the two species of akrasia suggests that he thinks they belong to one genus, so that one can be akratic by instantiating either one or the other. Equating akrasia with weakness of will thus gives us a *disjunctive account* of the latter: one shows weakness of will *either* by violating one's evaluative commitments (acting against one's best judgments) *or* by violating one's executive commitments (acting against one's resolutions). Here then we have a clear disagreement with Holton, who took weakness of will to consist just in the second disjunct.

To support this claim, Mele reports several empirical studies he conducted on the matter. The first two involve asking ordinary people (university students) what "weakness of will" means to them. In the first study, Mele asks his subjects to define what they mean by it. He reports that while only "eleven of the students (about 15%) mentioned doing something one knew or believed one should not do… only one student (about 1.4%) mentioned doing something one chose, decided, intended, or resolved not to do" (p. 396), In the second study, rather than giving them free rein, Mele asks his subjects to choose between three options:

A. Doing something you believed or knew you shouldn't do (for example, going to a party even though you believed it would be better to stay home and study).
B. Doing something you decided or intended not to do (for example, going to a party even though you decided to stay home and study).
C. Neither. The descriptions are equally accurate or inaccurate.

The results were that "49% gave the believed/knew response; 33% gave the decided/intended response; and 18% gave the third response" (p. 396). Mele contends that the results of these first two studies provide some evidence against Holton's resolution account and provide some evidence in favor of his own view.

This is effective as an *ad hominem* response to Holton, who himself made a claim about how ordinary people would gloss the idea of weakness of will. But let us pause to consider how much weight we should put on these two studies in elucidating the ordinary notion. In attempting to discover whether violating a resolution is more central to the ordinary notion of weakness of will, should we rely on the theoretical principles that ordinary people articulate in brief experimental conditions? The methodological trend among experimental philosophers has not

been to do so; they have appealed instead to people's application of a particular term or concept. This is because ordinary people are assumed to be rather good at recognizing when a certain ordinary concept (given that they possess it) applies in concrete cases, but not so good at recognizing the abstract principles that govern the application of these concepts. It is the role of the theorist to articulate the principles given the judgments of those who possess the concept. (Compare the practice of linguists, who place weight only on the concrete grammatical judgments of their subjects, not on their subjects' theories of grammar.) There may be something to be gained from examining how the folk articulate what they think does or does not count as displaying weakness of will. But we should put very little weight on such responses, especially when the issues are subtle.

Mele does move to more standard methods in his latter two studies. In Study 3, he asked subjects to make a judgment about a case in which judgment-violation and resolution-violation come apart. The scenario is adapted from one Mele has discussed in the past:

> Joe believes that it would be best to quit smoking cigarettes. He is thinking again—this time on New Year's Eve—about when to quit. He knows that quitting will be hard and unless he picks a good time to start he will fail. Joe judges that it would be best to smoke his last cigarette tonight and to be smoke free from then on. When he reports this to Jill, his wife, she asks whether this is his New Year's resolution. He says, "Not yet. I haven't yet actually decided to quit. Making that decision will be hard. To make it, I'll really have to psych myself up. I've been smoking for forty years. I believe I can quit, but I would definitely miss smoking." In the end, Joe fails to decide to quit smoking. Tomorrow, he smokes less than usual, but he has his first cigarette minutes after he awakes, as always. However, he could have decided to quit, and if he had he would have quit. (p. 401, n. 9)

Subjects were then asked to report on a Likert scale their degree of agreement or disagreement with the following claim: "Joe displays some weakness of will in this story." Given that in the story Joe acts against his best judgment but doesn't violate a resolution (since he doesn't ever intend to quit), Holton's hypothesis should make predictions about subjects' responses that differ from those of Mele's disjunctive view. Holton should predict that participants will tend to disagree with the claim while Mele should predict that they will tend to agree. Scoring "strongly agree" as 1 and "strongly disagree" as 7, the mean response was 2.68 (between "moderately agree" and "slightly agree"). As Mele reports, 80% agreed with the assertion (by providing a response of 1, 2, or 3), yet only 16% disagreed (providing a response of 5, 6, or 7). Mele contends that this counts against Holton's view: "this is evidence that an ordinary notion of weakness of will is such that Joe counts as displaying weakness of will—even though he does not act contrary to an intention. What he does act contrary to is his better judgment" (p. 402).

However, there are three serious worries here. First, there is a reasonable way to read the scenario so that it does contain a resolution-violation. After all, Joe appears to think he should decide to quit and that he is steeling himself for it. Joe says: "I haven't yet actually decided to quit. Making that decision will be hard. To make it,

I'll really have to psych myself up." It may appear to some readers that Joe has resolved to decide to quit (a sort of second-order intention): he is forming an intention to make the official decision to plan on quitting. And this appears to be a resolution, as opposed to just an intention, because he is clearly anticipating that he will be tempted not to stick to his second-order intention. As he says, the decision will be hard and he'll have to "psych himself up." This sounds much more like the smoker who has a first-order resolution to quit than the merely akratic smoker who judges quitting to be the best option but just isn't moved to do what is best. If this is right, a significant number of Mele's subjects may have agreed that Joe displays weakness of will in the vignette, but only because they were picking up on the reading on which he is violating a resolution, albeit a second-order one. If Mele had asked subjects whether they agreed with the claim that Joe displays weakness of will specifically in failing to quit smoking, this wouldn't be much of a problem. However, he asked them about their agreement with a more general claim ("Joe displays some weakness of will in this story") which can apply to the relevant second-order resolution.

The second problem concerns the wording of the question that Mele asked. Subjects were asked whether Joe showed *some* weakness of will. That might be an effective question if one were only concerned with refuting the letter of Holton's earlier account. Suppose though that Holton was right in thinking that the core issue in weakness of will is resolution-violation; judgment-violation has some role to play, but a comparatively minor one. That would be in the spirit of Holton's account. In such a case, though, subjects might still accept that Joe showed some weakness of will; the claim is, after all, a very weak one.

The final problem with Mele's third study is that the result differs markedly from a similar one he conducted ("Study 3a" n. 10, p. 402). Here Mele varied the dependent measure, asking participants to report their degree of agreement with a slightly different claim: "Joe does not display any weakness of will in this story." Now, this is the negation of Mele's original claim, so one would expect to get a result that is the mirror image of the earlier one: disagreement in Study 3 should be matched by agreement in Study 3a. However, in this subsequent study disagreement had fallen to 58% (compared to 80% agreement prior), whilst agreement was up to 38% (compared to 16% disagreement prior). As Mele notes, this result is disconcerting. Still, he takes comfort in the fact that a majority in each study (80% in one, 58% in the other) provided a response that is *not* in accord with Holton's view, even though Study 3a did not yield a strong majority. But a discrepancy of this magnitude should make us worry that something is going wrong.

Realizing the uncertainty surrounding these studies, Mele conducted one more (Study 4) in an attempt to replicate the results of Study 3 while using a slightly different way of measuring responses. Using the same scenario involving Joe above, Mele asked subjects to respond with "Yes" or "No" to the question: "Does Joe display any weakness of will in this story?" A large majority (73%) answered in the affirmative, seemingly providing more evidence against Holton's view. But the first two worries about Study 3 apply here as well. Given that the story is the same, some participants may have read Joe as violating a second-order resolution. Furthermore, in answering "Yes," they are committed to a very weak claim—namely, that Joe

displays *some* weakness of will. It's not surprising that a large majority of people opted for saying "Yes," especially given such a forced, dichotomous choice.

While Mele concludes that Holton's view is discredited and his disjunctive account is well-supported, the worries we have raised here call out for further investigation. So we conducted three experiments in an attempt to provide more conclusive results.

3 Experiment 1: Varied vignettes

In our first experiment, we developed a factorial design to look for effects of either of the relevant variables—Judgment-Violation (JV) or Resolution-Violation (RV)—on pre-theoretical judgments about cases potentially involving weakness of will. The 97 subjects from around the University of California, Santa Barbara were randomly assigned to be in one of the four conditions, yielding about 25 participants in each.[6]

In the first condition, subjects were presented with a vignette in which an agent (Newman) performs an action (eating donuts) that violates his judgment and resolution.

> *Newman's Diet* (JV, RV):
> Newman is worried about his weight. His doctor has told him he needs to lose weight or he'll likely die of a heart attack in the near future. In light of this, Newman thinks it's best to go on a diet and plans to do so. He stocks up on healthier foods and buys a book on how to lose weight.
>
> It's two weeks into Newman's diet, and he's at work chewing on a carrot. However, one of his co-workers brings in a large box of fresh donuts from the local bakery, and Newman loves donuts; they're his favorite. Even though he has gone two whole weeks without eating any unhealthy food, like these donuts, Newman succumbs to temptation and eats one each of his favorites. Afterwards, he is riddled with guilt for going against what he thought was best and had planned not to do.

Participants were then asked to indicate (on a Likert scale) their degree of agreement or disagreement with the following claim: "Newman displays weakness of will in eating the donuts." Here we avoided using quantificational terms—such as "some" or "any"—to avoid complications with weaker versus stronger claims as in Mele's studies. We also made the claim specifically about weakness of will with respect to the action in question, rather than in the story in general, to avoid the first problem noted with Mele's design.

[6] Subjects were predominantly between the ages of 18–24. Some were approached in a classroom setting using paper surveys while others were solicited via email to participate in a web-based survey. The vast majority were undergraduates at UC Santa Barbara. For the web-based version, appropriate measures were taken to avoid automatic responses from robots, web crawlers, etc.

Subjects in the second condition were presented with a scenario in which Christabel performs an action (adultery) that is a violation of her judgment but not any resolution.

> *Christabel's Affair* (JV, ∼RV):
> Christabel, a happily married Victorian lady, is tempted to start an extra-marital affair with William, a man she has recently met.
>
> She knows it is likely to be disastrous. People are bound to find out, and it will ruin her marriage and her reputation. Moreover, she considers it morally wrong. So she thinks it's not the best option. Nevertheless, she is not moved by these considerations, and has planned to go ahead with it anyway.
>
> When the weekend comes, Christabel follows through with her plan: she sneaks out late at night, meets William, and they start an affair.

Participants were then asked to indicate their degree of agreement or disagreement with the following claim: "Christabel displays weakness of will in having the affair."

For the third condition, subjects read a vignette in which the protagonist (Rocky) performs an action that is consistent with his evaluative judgment but violates a resolution of his.

> *Rocky's Loss of Nerve* (∼JV, RV):
> Rocky, who has promised his mother that he would never play tackle football, has just been invited by some older boys to play in tomorrow's game. Given his promise to his mother, he thinks it would be best not to play. But he really wants to, so he decides to play anyway.
>
> However, when the time comes, Rocky suffers a failure of nerve. He doesn't show up for the game—not because he thinks it best not to play, but because he's afraid. He wouldn't have played even if he had thought it best to do so.

Respondents were then asked to indicate their degree of agreement or disagreement with the following claim: "Rocky displays weakness of will in not showing up for the game."

Subjects in the fourth condition were presented with a story in which a woman (Kima) performs an action (adultery) that is neither a judgment- nor resolution-violation.

> *Kima's Affair* (∼JV, ∼RV):
> Kima is working late at the office with her co-worker, Omar. As they joke together about their relentless boss, she realizes she is greatly attracted to Omar—both physically and intellectually.
>
> Although Kima is married and her husband is good to her, she doesn't much care about his feelings. She thinks it would be best to just go ahead and seduce Omar into having sexual intercourse. So she walks into Omar's office and carries out her plan to seduce him. Omar doesn't take much persuading, and they proceed with the affair.

The participants were then asked to indicate their degree of agreement or disagreement with the following claim: "Kima displays weakness of will in having the affair."

Given this design, we can construe Mele's and Holton's views as competing hypotheses. The first and last conditions are not in dispute; both hypotheses predict relatively high levels of agreement in the first condition and disagreement on average in the fourth condition. Despite this common ground on these cases, including them will allow us to acquire a richer set of data and to compare any statistically significant effects of the independent variables. What about the second and third conditions? Mele's disjunctive hypothesis should predict no significant difference between them and that they will yield on average relatively high levels of agreement with the relevant assertion (levels that near that of Condition 1). On the other hand, Holton's resolution hypothesis predicts at least that agreement will be significantly higher in Condition 3 than in Condition 2 (given that the former involves a resolution-violation while the latter doesn't).

As a first pass at examining the results, we can consider the percentage of subjects for each condition who either agreed, disagreed, or neither agreed nor disagreed with the relevant statement. We scored responses in the reverse of the order in which Mele did so that an increase in the number corresponds more intuitively to an increase in agreement (so 7 is strongly *agree* while 1 is strongly *disagree*). Grouping responses into the three categories of agree, disagree, or neither, we found that a majority of participants (74%) in the first condition (JV, RV) *agreed* that Newman displays weakness of will (by providing a response of either 5, 6, or 7). This is precisely what we all should expect. Similarly, a majority (63%) of subjects in the last condition ($\sim$JV, $\sim$RV) *disagreed* with the claim that Kima displays weakness of will (by providing a response of 1, 2, or 3).[7] The results for the middle two cells indicate much weaker trends. In Condition 2 (JV, $\sim$RV), 50% of subjects agreed (to some extend or other) that Christabel displays weakness of will, but 33% disagreed and 17% were ambivalent. Similarly, in Condition 3 ($\sim$JV, RV), 50% agreed that Rocky displays weakness of will, but 27% disagreed and 23% were ambivalent.

Examining the mean score of responses in each condition, we find a similar trend. Table 1 and Fig. 1 display means for each cell.

A statistical analysis of these data shows that there were two separate significant effects for each variable.[8] Participants tended to give higher ratings of agreement

[7] It may seem quite odd that there isn't a strong majority of participants providing the expected response here, as there is in the first condition. But there are several plausible explanations of this, which aren't mutually exclusive. First, given that Kima does something many would likely consider immoral, some subjects may have felt compelled to agree that Kima displays weakness of will simply to allocate the stigma attached to it. Second, the expected response to Kima's case involves rejecting the statement presented, and there is some psychological evidence that we are more inclined to accept a proposition, at least initially, than to reject it (Gilbert 1991). While this may explain the minor deviations here, it doesn't seem able to explain more drastic ones, such as those we observe in Mele's Study 3a.

[8] The data were subjected to a 2 (JV vs. $\sim$JV) $\times$ 2 (RV vs. $\sim$RV) between-subjects analysis of variance (ANOVA). There was a main effect of Judgment-Violation, $F(1, 93) = 10.4$, $p < .01$, and a main effect of Resolution-Violation, $F(1, 93) = 8.9$, $p < .01$. There was no significant interaction effect. (Special thanks to Joshua Knobe for assistance here.)

Table 1 Mean responses for Experiment 1

	RV	~RV
JV	5.57	4.29
~JV	4.19	3.08

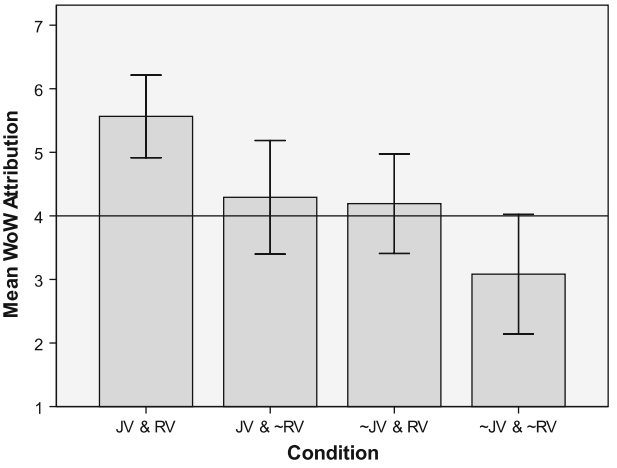

Fig. 1 Mean responses for Experiment 1

when either variable was present. So these results certainly do count against Holton's resolution account given that both variables had a significant effect on whether people thought a case involved weakness of will. This may appear initially to show that the ordinary conception conforms to Mele's disjunctive view—it involves either a judgment-violation *or* a resolution-violation. However, a closer look at the means indicates something more complex is going on.

As both hypotheses predict, and as we all should expect, there is relatively high agreement with the attribution of weakness of will (or the exhibition of it) in Newman's case involving both kinds of violation. Similarly, it is no surprise that we find disagreement on average in the final condition in which Kima doesn't violate her judgment or resolution at all. But the key results are in the middle two conditions, and they may seem to support Mele's hypothesis, especially since there is no significant difference between these means.[9] However, Mele should also predict that the average level of agreement in the middle conditions would be relatively high, at least close to the mean of Condition 1 (RV, JV). After all, if the disjunctive account is true (i.e. if cases involving either kind of violation are sufficient for exhibition of weakness of will according to ordinary folks), then we would expect competent speakers to tend to *agree* with the relevant attribution. Yet cases involving only judgment-violation or only resolution-violation produced means very near the midpoint (neither agree nor disagree). So our subjects tended to be neutral with respect to such cases. Likewise, a purely *conjunctive* view appears to

[9] Confirmed by an independent samples *t*-test, $t(48) = .173$, $p = .86$.

have a similar problem accounting for these results. It should predict that the middle two conditions would yield means much closer to that of Condition 4 than we found.

This indicates that neither Holton nor Mele were quite on the right track. When both forms of violation are *absent*, our subjects tended to think the protagonist doesn't display weakness of will. But when both variables are *present*, they tended to think the protagonist does display weakness of will. Given the ambivalence produced in the middle two conditions, this suggests that while both variables may be necessary for full, confident application of the concept, neither alone is sufficient. Perhaps then we should think of the ordinary concept of weakness of will as a proto-type or cluster concept (Rosch 1975). Contra both theorists, there doesn't appear to be a simple notion here with necessary and sufficient conditions for its application—disjunctive or otherwise. Rather, each variable plays *contributory* roles in the application of the concept of weakness of will. Each counts to some extent toward application of the concept, but neither is sufficient on its own.[10] We don't want to commit ourselves to a general proto-type theory of *all* concepts; but these data do provide some evidence that the ordinary notion of weakness of will is operating this way.

One might object at this point that our vignettes in this experiment are not uniform enough to isolate our two variables.[11] After all, the vignettes do clearly vary in topic (dieting, keeping a promise, and adultery). The various moral, evaluative, and normative differences here may be cause for alarm. To address this worry, we ran another experiment.

4 Experiment 2: Uniform vignettes

Our second experiment is the same in form as our first, differing only in some key respects. First, we made the vignettes more uniform to see whether our previous results could be replicated. We constructed four cases involving Carl, who has always wanted to go skydiving, but who thinks, on the advice of his physician, that it's best he doesn't. The vignettes differ on whether Carl resolves to go or resolves not to, and on whether he ends up jumping. In the first he resolves not to go, and then does jump (JV, RV); in the second he resolves to go and then does jump (JV, ∼RV); in the third he resolves to go and then doesn't jump (∼JV, RV) and in the fourth he resolves not to go and doesn't jump (∼JV, ∼RV). The full vignettes are included in the Appendix. Here we not only made the cases as uniform as we could without sacrificing natural-sounding stories, we also opted for a morally-neutral action (sky diving) to ward off worries about the moral valence of the case having an independent effect on subjects' responses. Second, we attained a sample size that was significantly larger (n = 274, about 68 responses per condition) and more diverse (undergraduate students from a variety of disciplines and three different universities).

To ward off the uniformity objection to Experiment 1, and thus retain the support for our proto-type account, we would ideally find the exact same results in

[10] See Knobe and Preston-Roedder (2009) for similar results on the concept of *valuing*.

[11] Thanks especially to Craig Roxborough for originally pressing us on this issue.

Table 2 Mean responses for Experiment 2

	RV	∼RV
JV	4.01	3.00
∼JV	3.54	2.00

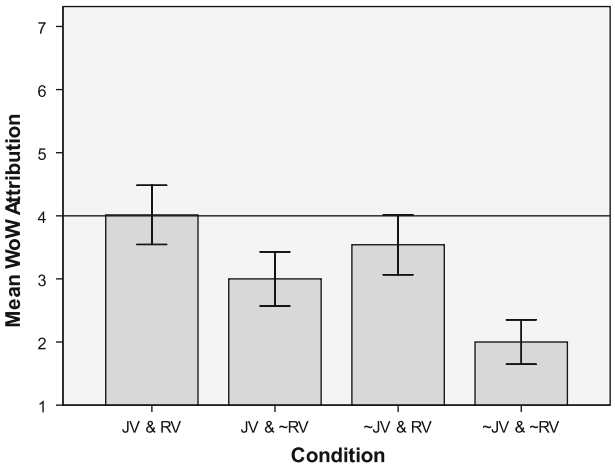

Fig. 2 Mean responses for Experiment 2

Experiment 2 as we found in Experiment 1. However, that isn't exactly the case. We did again find that subjects were more likely to agree that Carl displayed weakness of will when there was both resolution- and judgment-violation, less likely to when neither of these factors held, and in between when just one held. But there was a substantial difference from the results of the first experiment. Table 2 and Fig. 2 show the mean score for responses in the four conditions.

Figure 2 shows that the means for each condition in this experiment are about the same *relative* to one another as they were in the first. But for each condition they went down uniformly by about one point on the scale. This is fairly puzzling at first because we should expect most people to at least agree that the first vignette (JV, RV) involves weakness of will. But the mean is around 4.0 (neither agree nor disagree). The most common score is 6 (moderately agree) but the mean is pulled down to the midpoint by a sizable group of subjects providing a response of 2 (moderately disagree). So it looks as though there is a sub-group who are reading the case in such a way that they don't tend to think Carl displays weakness of will in jumping, even though he thought it was best not to and resolved not to do it.

Likewise, a statistical analysis of the data shows that the two variables had a significant effect on subjects' responses.[12] So, again, contra Holton's resolution

[12] The data were subjected to a 2 (JV vs. ∼JV) × 2 (RV vs. ∼RV) between-subjects ANOVA. There was a main effect of Judgment-Violation, $F(1, 270) = 11.4$, $p < .01$, and a main effect of Resolution-Violation, $F(1, 270) = 34.3$, $p < .01$. There was no significant interaction effect.

hypothesis, both variables are significantly affecting people's attributions of weakness of will. These data may seem to lend some support to the resolution account, given that the mean response for the case involving a resolution-violation but no judgment-violation is higher than the mean response for the case involving judgment-violation but no resolution-violation (3.54 vs. 3.00). However, these differences are not significant.[13]

What could explain these rather odd results? We might suspect this shows that different people were picking up on different features of the cases, and so these data don't reflect anything about the ordinary conception of weakness of will. But many subjects who filled out the optional portion of the survey asking for an explanation of their response did seem to be picking up on the relevant notion of weakness of will. For example, in the fourth condition ($\sim$JV, $\sim$RV), a large number of participants disagreed with the claim that Carl displayed weakness of will, as we would expect. And those who provided explanations for that choice very frequently referenced something like Carl's sticking to what he chose to do, not succumbing to temptation, and so on. Another factor that may have played a role in driving down agreement is the fact that Carl is exhibiting *courage* in jumping out of a plane, which to some may have seemed odd to describe as a case of weakness.[14] This alone, though, wouldn't uniformly explain the drop in agreement since in two of the four cases Carl doesn't actually jump (and so doesn't exhibit courage). But it certainly could be playing some role as well. We suspected, then, that perhaps the difference in results is at least partly due to our opting for morally-neutral cases. To test this explanation, we ran another experiment.

5 Experiment 3: Valence

In our final study, we wanted to test whether the moral valence of the case affected subjects' attributions of weakness of will. Since we wanted to explain the results of Experiment 2, we needed to determine which aspects of Carl's case could have been morally infused. Ultimately, what Carl intended to do and what he ended up doing in the cases of succumbing (either going or refraining from sky diving) were fairly morally neutral—or, more broadly, normatively and evaluatively neutral.[15] So we set out to look for effects of the moral valence of either the intention or the action.

To this end, we developed another factorial design with two variables: Action-Valence (either neutral or bad) and Intention-Valence (either neutral or bad). We developed four uniform vignettes accordingly. Each one involved Phil, who is either resolving to read some French literature (neutral) or some Nazi literature (bad) and succumbing to either go with his friends to drunkenly bully people (bad) or watch a

[13] Confirmed by an independent samples *t*-test, $t(134.7) = -1.69$, $p = .094$.

[14] Thanks especially to Al Mele for raising this potential explanation.

[15] In *some* cases, Carl did go against the advice of his physician by skydiving (though perhaps not intentionally under that description). And this is perhaps a violation of a norm. (Jonathan Way helpfully raised this issue.) But we submit that this is at least much *less* normatively infused than the relevant scenarios in our third experiment.

movie (neutral). In each case, Phil judges some course of action best, resolves to do it, but eventually succumbs. Here are the four, quite uniform, and rather concise cases (emphasis has been added here in order to make the differences more explicit):

Case 1: Neutral Intention and Neutral Action (NI, NA)
Phil has recently joined a *French class*. He is deeply committed to the class and to what they are trying to achieve, though he tends to put off reading the classic French texts that the *teacher* insists they all study. He finds them a bit of a drag.
This evening he has resolved to stay home and read some of the texts. But some friends call up and try to persuade him to come out with them. If things go as normal they'll *have a pizza and watch a movie*. He thinks it would be better to stay home and read as planned, but he gives in and goes with them.

Case 2: Neutral Intention and Bad Action (NI, BA)
Phil has recently joined a *French class*. He is deeply committed to the class and to what they are trying to achieve, though he tends to put off reading the classic *French texts* that the *teacher* insists they all study. He finds them a bit of a drag.
This evening he has resolved to stay home and read some of the texts. But some friends call up and try to persuade him to come out with them. If things go as normal they'll *hang out at the mall, have rather too many beers, and pick fights with some of the local immigrant kids*. He thinks it would be better to stay home and read as planned, but he gives in and goes with them.

Case 3: Bad Intention and Neutral Action (BI, NA)
Phil has recently joined a *Neo-Nazi group*. He is deeply committed to the group and to what they are trying to achieve, though he tends to put off reading the classic *Nazi texts* that the *group leader* insists they all study. He finds them a bit of a drag.
This evening he has resolved to stay home and read some of the texts. But some friends call up and try to persuade him to come out with them. If things go as normal they'll *have a pizza and watch a movie*. He thinks it would be better to stay home and read as planned, but he gives in and goes with them.

Case 4: Bad Intention and Bad Action (BI, BA)
Phil has recently joined a *Neo-Nazi group*. He is deeply committed to the group and to what they are trying to achieve, though he tends to put off reading the classic *Nazi texts* that the *group leader* insists they all study. He finds them a bit of a drag.
This evening he has resolved to stay home and read some of the texts. But some friends call up and try to persuade him to come out with them. If things go as normal they'll *hang out at the mall, have rather too many beers, and pick fights with some of the local immigrant kids*. He thinks it would be better to stay home and read as planned, but he gives in and goes with them.

Each vignette was randomly assigned to one of 117 undergraduate students in a Critical Thinking course at the University of California, Santa Barbara, yielding about 30 subjects per condition. After reading the vignette, participants recorded

Table 3 Mean responses for Experiment 3

Intention-Valence (I)	Action-Valence (A)	
	Neutral (NA)	Bad (BA)
Neutral (NI)	5.23	5.73
Bad (BI)	4.89	5.52

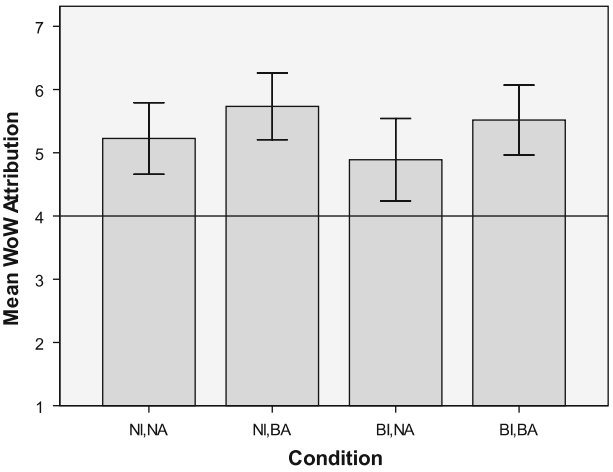

Fig. 3 Mean responses for Experiment 3

their degree of agreement or disagreement with the following claim: "Phil displays weakness of will in going with his friends." Table 3 and Fig. 3 show the mean score for each condition using the same Likert scale as in the previous experiments.

As Fig. 3 demonstrates, the mean response for each condition is, as we would expect, above the midpoint. So subjects on the whole tended to agree in each case that Phil displays weakness of will. Moreover, an analysis of the data revealed a significant effect of the Action-Valence variable such that subjects are more inclined to agree with the attribution of weakness of will when the valence of the action the agent ends up succumbing to perform is bad as opposed to neutral.[16] Surprisingly, there was no corresponding effect found for the Intention-Valence variable and there was no interaction effect. Figure 3 displays this well. The responses were higher in the conditions in which the action was bad (i.e. the second and fourth cases). When the valence of the action is neutral, the mean response is around 5 (slightly agree). When the valence of the action becomes bad, the mean jumps to around 6 (moderately agree).

We think these results help explain those of Experiment 2. They provide an explanation of why we found such low levels of agreement, even though we achieved the same spread, so to speak, from Experiment 1. This, we submit, provides a substantial case against the objection to Experiment 1 based on lack of

[16] $F(1, 113) = 4.10, p = .045$. Confirmed by a 2 (bad intention versus neutral intention) $\times$ 2 (bad action versus neutral action) between-subjects ANOVA.

uniformity. Experiment 3 suggests that at least one key reason we didn't find higher levels of agreement in Experiment 2 is due to our use of a more morally neutral case. Presumably a more morally infused case would bump up the spread, which would replicate the results of Experiment 1, and so help to support our initial proto-type account of the ordinary concept of weakness of will. Of course, some of the cases in Experiment 1 didn't exactly have a moral valence—it seems a bit of a stretch to think of a failure to maintain a diet as a *moral* failure. But the actions they performed did all end up violating some clear norm or amount to something we'd expect the average person to consider bad, though not necessarily morally bad. And that's the general kind of valence we're concerned with—i.e. normative/evaluative, broadly speaking, which is what many have found as the general factor in phenomena like the Knobe effect (see e.g. Pettit and Knobe 2009). That is, the kinds of factors that appear to have such a pervasive impact on so much of our thinking isn't specific to morality, but rather to violating norms more generally.[17]

In addition to supporting our proto-type account, the results of Experiment 3 are independently interesting. They indicate that normative or evaluative considerations affect people's judgments about whether an agent is being weak-willed. We're not sure what to make of this additional finding. While odd in certain respects, there is a wealth of data indicating that normative and evaluative considerations affect our application of various notions (Pettit and Knobe 2009).

6 Conclusion

The results of our first experiment provide some evidence that neither Holton's resolution account nor Mele's disjunctive account were quite correct. Instead, a proto-type account of the ordinary concept of weakness of will seems to best explain the data. After all, in Experiment 1, both types of violation were required for the mean response to be well above the midpoint. And the means for the two cases in which only one type of violation was present (judgment or resolution), were near the midpoint of "neither agree nor disagree." Moreover, the proto-type account is consistent with the results of Experiment 3. In all four vignettes, participants tended to agree that weakness of will was displayed by the protagonist, but both types of violation were also present.

However, Experiment 1 is open to the criticism that our vignettes didn't isolate the two variables at issue. Experiment 2 was developed in an attempt to address this problem. Though we failed to find the *exact* same results as in our first experiment, we hypothesized that this was due to the lack of significantly normative or evaluative valence of the cases. Experiment 3 provided some confirmation of this hypothesis. Experiment 2 was meant primarily to ward off an objection to our conclusion from Experiment 1. It didn't fully, but we think supplementing it with the results of Experiment 3 does, and this protects our initial tentative conclusion from Experiment 1. However, Experiment 3 also has the independently interesting

[17] For an attempt to explain the Knobe effect along these lines, see Holton (2010).

finding that the valence of the action, but not the intention, does affect at least people's confidence in their attributions of weakness of will.[18]

So what should we conclude from this? What in the world is the ordinary concept of weakness of will? Holton indeed should be much less confident in the existence of an ordinary notion of weakness of will that only involves resolution-violations. Furthermore, our results indicate that focusing on either resolution-violations or judgment-violations (or both) exclusively isn't quite right. The normative valence of the action seems to play a role just as the other two variables do and in a contributory rather than a classical way. While we shouldn't consider the matter closed based on a few experiments, a proto-type account does provide a good explanation of the current data.

Of course, one might object to our entire project here on the grounds that looking for an ordinary concept of weakness of will is rather dubious or even perverse.[19] While we're sympathetic to this worry, we have two things to say in response. First, regardless of whether we should be concerned as theorists about how ordinary folks use the term "weakness of will," philosophers have often either explicitly or implicitly had a keen interest in the ordinary notion by considering judgments about hypothetical cases.[20] Second, our data do seem to indicate that there is a real notion here. After all, significant majorities of people in our studies clustered around agreement or disagreement with the attribution of weakness of will depending on some of the very factors that we would expect. If there were no real ordinary notion, we would expect much more erratic and puzzling data.

Finally, what does this mean for theorists interested in weakness of will? Does it matter whether philosophers employing the phrase are theorizing about something that has straightforward necessary and sufficient conditions for its application and is independent of normative or evaluative considerations? It depends on the theorist. Some clearly take "weakness of will" to be a term of art which picks out a certain phenomenon they're interested in, such as judgment-violation. Davidson (1970) may be an example. Others, however, are concerned with the folk notion of weakness of will. Either way, perhaps we can at least conclude this: we should be clear about whether we are interested in the ordinary notion of weakness of will, or just judgment-violations, or just resolution-violations, or something else entirely.

Acknowledgments This paper has benefitted from numerous comments and discussion. We thank in particular: Toby Handfield, Josh Knobe, Al Mele, Craig Roxborough, Steve Stich, Jesse Summers, Jonny Way, Jonathan Weinberg, and an anonymous referee for this journal. Versions of this paper were presented by May at the 2010 Pacific Division meeting of the APA in San Francisco, the 2010 Eastern Division meeting of the APA in Boston, a conference at the University at Buffalo, the Philosophical Society at SUNY Fredonia, and a graduate seminar of the Mind/Brain/Behavior group at Harvard University. The paper was also discussed on the weblog Flickers of Freedom. The audiences and

[18] We say "their confidence" instead of "their judgments" only because the valence didn't appear to drive participants from denying weakness of will to attributing it; it only moved them up on the same agreement side of the midpoint. See May et al. (2010) for a similar approach to differences on one side of a Likert scale, though on the topic of knowledge attributions.

[19] Thanks especially to Stephen Stich for pressing this concern.

[20] Just focusing on recent work, one can see this in, for example, Dodd (2009), Cohen and Handfield (forthcoming), Levy (forthcoming), and of course Holton and Mele.

commentators on every occasion provided valuable input that we greatly appreciate. A word on the authorship of this paper: May did the empirical work; Holton did the criticism of Holton; the rest is joint.

Appendix: Vignettes for Experiment 2

The differences between these cases are italicized to help the reader see them. But, of course, the vignettes given to subjects did not contain any emphasis.

Vignette 1 (JV, RV)

For as long as he can remember, Carl has wanted to have a go at skydiving. A couple of colleagues at work are also keen, and together they discuss signing up for a course. He is excited about the possibility, and understandably a little anxious. However, when he mentions the possibility to his doctor, he is advised, given his medical history, not to do it. He is very disappointed, but thinking it over reluctantly concludes that his doctor is right: the best thing would be to let his friends go without him.

Still, in Carl's mind that doesn't close the matter. He still finds the idea of jumping terribly exciting. His doctor hasn't actually forbidden him to go, and he doesn't need a medical certificate from him. When his friends encourage him to sign up for the course, he seriously considers doing so, even though he still thinks that, given the health concerns, it would be best not to. Finally he *resolves not to go*.

When the first day of the course comes, Carl's *friends call him to say that there is an empty place and urge him to come*. Despite his earlier resolution, Carl *gives in and goes with them*. He completes the preliminary training, and makes his first jump.

Please indicate your degree of agreement or disagreement with the following statement: "Carl displays weakness of will in jumping."

Vignette 2 (JV, $\sim$RV)

For as long as he can remember, Carl has wanted to have a go at skydiving. A couple of colleagues at work are also keen, and together they discuss signing up for a course. He is excited about the possibility, and understandably a little anxious. However, when he mentions the possibility to his doctor, he is advised, given his medical history, not to do it. He is very disappointed, but thinking it over reluctantly concludes that his doctor is right: the best thing would be to let his friends go without him.

Still, in Carl's mind that doesn't close the matter. He still finds the idea of jumping terribly exciting. His doctor hasn't actually forbidden him to go, and he doesn't need a medical certificate from him. When his friends encourage him to sign

up for the course, he seriously considers doing so, even though he still thinks that, given the health concerns, it would be best not to. Finally he *resolves to sign up anyway*.

When the first day of the course comes, *Carl gets increasingly anxious about the jump*. As he goes through the preliminary training, he *starts to wonder whether he will have the nerve to jump*, but he *repeats to himself his resolution to go through with it*. When the time of the jump finally arrives, *he is terrified*. Nevertheless, *he manages to jump*.

Please indicate your degree of agreement or disagreement with the following statement: "Carl displays weakness of will in jumping."

Vignette 3 ($\sim$JV, RV)

For as long as he can remember, Carl has wanted to have a go at skydiving. A couple of colleagues at work are also keen, and together they discuss signing up for a course. He is excited about the possibility, and understandably a little anxious. However, when he mentions the possibility to his doctor, he is advised, given his medical history, not to do it. He is very disappointed, but thinking it over reluctantly concludes that his doctor is right: the best thing would be to let his friends go without him.

Still, in Carl's mind that doesn't close the matter. He still finds the idea of jumping terribly exciting. His doctor hasn't actually forbidden him to go, and he doesn't need a medical certificate from him. When his friends encourage him to sign up for the course, he seriously considers doing so, even though he still thinks that, given the health concerns, it would be best not to. Finally he *resolves to sign up anyway*.

When the first day of the course comes, Carl gets increasingly anxious about the jump. As he goes through the preliminary training, he starts to wonder whether he will have the nerve to jump, but he repeats to himself his resolution to go through with it. When the time of the jump finally arrives, he is terrified. He finds that he is just too scared to jump. He remains in the plane, and returns to the airfield.

Please indicate your degree of agreement or disagreement with the following statement: "Carl displays weakness of will in *not* jumping."

Vignette 4 ($\sim$JV, $\sim$RV)

For as long as he can remember, Carl has wanted to have a go at skydiving. A couple of colleagues at work are also keen, and together they discuss signing up for a course. He is excited about the possibility, and understandably a little anxious. However, when he mentions the possibility to his doctor, he is advised, given his medical history, not to do it. He is very disappointed, but thinking it over reluctantly concludes that his doctor is right: the best thing would be to let his friends go without him.

Still, in Carl's mind that doesn't close the matter. He still finds the idea of jumping terribly exciting. His doctor hasn't actually forbidden him to go, and he doesn't need a medical certificate from him. When his friends encourage him to sign

up for the course, he seriously considers doing so, even though he still thinks that, given the health concerns, it would be best not to. Finally he *resolves not to go*.

When the first day of the course comes, Carl's friends call him to say that there is an empty place and urge him to come. But *Carl is firm*; he *remains at home* while his friends jump.

Please indicate your degree of agreement or disagreement with the following statement: "Carl displays weakness of will in *not* jumping."

References

Cohen, D. & Handfield, T. (forthcoming). Rational capacities, resolve, and weakness of will. *Mind*.

Davidson, D. (1970). How is weakness of the will possible? In J. Feinberg (Ed.), *Moral concepts*. Oxford University Press. (Reprinted in as ch. 2 in his *Essays on Actions and Events*, 1980/2001, Oxford: Clarendon Press.)

Dodd, D. (2009). Weakness of will as intention-violation. *European Journal of Philosophy, 17*(1), 45–59.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46*(2), 107–119.

Holton, R. (1999). Intention and weakness of will. *Journal of Philosophy, 96*(5), 241–262.

Holton, R. (2009). *Willing, wanting, waiting*. Oxford: Clarendon Press.

Holton, R. (2010). Norms and the Knobe effect. *Analysis, 70*(3), 1–8.

Knobe, J., & Preston-Roedder, E. (2009). The ordinary concept of valuing. *Philosophical Issues, 19*(1), 131–147.

Levy, N. (forthcoming). Resisting weakness of the will. *Philosophy & Phenomenological Research*.

May, J., Sinnott-Armstrong, W., Hull, J., & Zimmerman, A. (2010). Practical interests, relevant alternatives, and knowledge attributions: An empirical study. *Review of Philosophy and Psychology*, special issue on Psychology and Experimental Philosophy, E. Machery, T. Lombrozo, & J. Knobe (Eds.), *1*(2), 265–273.

Mele, A. (2010). Weakness of will and akrasia. *Philosophical Studies, 150*(3), 391–404.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language, 24*(5), 586–604.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192–233.