Ethic Theory Moral Prac (2017) 20:519–536
DOI 10.1007/s10677-017-9810-5

Challenging Doris' Attack on Aggregation: Why We are Not Left "Completely in the Dark" about Global Virtues

William Fleeson[1] · **Eranda Jayawickreme**[1]

Accepted: 19 April 2017 / Published online: 6 May 2017

Abstract Aggregation (the process of collecting multiple observations of behavior and averaging them before predicting behavior) shows that virtue-relevant behavior is indeed highly predictable, and that individual differences in global virtues do indeed exist. Aggregation is a key response to the situationist argument against the existence of broad virtues. However, a concern with aggregation is that, because it is an average, the specifics of what are included in that average matter. In particular, if heinous actions could be included in the average, then aggregates cannot provide enough confidence that the holders of high aggregates (e.g., highly compassionate people) have not conducted heinous actions and thus cannot provide enough confidence that such people qualify as virtuous. Doris (2002) has challenged aggregation with this concern, and no one has responded substantively to this challenge. If Doris' challenge is in fact correct, then the situationist argument against the existence of broad virtues stands. In the article, we present a full response to this concern. We argue that aggregation does not in fact allow heinous exceptions, because aggregates do indeed predict extreme single behaviors very well. In fact, aggregates do allow confidence that holders of high aggregates do not commit heinous actions. Thus, Doris' rejection of the aggregation solution does not defeat aggregation, aggregation continues to stand in the defense of global virtues, and the situationist argument does not threaten the existence and predictive power of global virtues. Models of traits that rely on aggregates, such as Whole Trait Theory (Fleeson and Jayawickreme 2015), may provide useful post-situationist models of virtues.

Keywords Situationism · Aggregation · Morally exceptional · Virtue · Traits

✉ William Fleeson
fleesonw@wfu.edu

Eranda Jayawickreme
jayawide@wfu.edu

[1] Wake Forest University, Winston Salem, NC, USA

The purpose of this paper is to defend one of the principal arguments that establishes empirical support for the existence and importance of global character traits. Specifically, one major line of defense of global traits is the principle of aggregation– the sum or average of a set of multiple measurements is more stable and unbiased than any single measurement from that set. Without this principle, the belief in broad traits becomes more difficult to reconcile with the empirical evidence demonstrating behavioral variability and inconsistency. Such behavioral variability and inconsistency arguably demonstrates that traits have little to no meaningful impact on behavior and thus render any talk of virtue and character fruitless (Doris 1998; Doris 2002; Alfano 2013; Harman 1999; Vranas 2009). However, aggregation reveals massive consistency and predictability of behavior, suggesting that scientific evidence provides compelling support for the philosophical and psychological study of virtue (e.g., Jayawickreme et al. 2014; Slingerland 2011).

Most philosophers sympathetic to the situationist position do not address this important aggregation defense of broad character traits in their work. Doris (2002) is one notable exception. In his book *Lack of Character* (Doris 2002), he lays out strong objections to aggregation as a defense of broad traits. As far as we know, there has been no attempt to respond to these objections to the principle of aggregation. This is important, because if aggregation is to be a major defense of virtues, then Doris' objections to aggregation have to be answered.

We believe Doris' objections have not been answered both because the objections rely on a compelling concern, and because aggregation is complex. The concern is that aggregates may not be a sufficient basis for designating virtuous individuals, because aggregates may create high consistency precisely by masking heinous actions committed by those individuals designated as virtuous by the aggregates. In this paper, we use simulated data to critically examine this concern, and show that the concern is ultimately unfounded. This concern and Doris' arguments do not in fact pose a significant threat to the principle of aggregation; thus, defenses of global broad traits that rely on aggregation are not in fact undermined, the aggregation defense of broad global traits continues to hold off the situationist critique, and aggregates can be the basis of designating individuals' levels of virtue. We propose that models of traits that rely on aggregates, such as Whole Trait Theory (Fleeson and Jayawickreme 2015), may provide useful post-situationist models of virtues.

1 Overview of the Situationist Critique of Global Traits and Virtues, and the Counter-Argument

Our purpose in this paper is very specifically to defend aggregation as a strong response to the situationist critique. We wish to defend it against an important concern about aggregation, articulated in Doris' counterargument to aggregation as a strong response to the situationist critique. However, to understand why aggregation is a strong response, and to be in a position to evaluate Doris' counter-argument against aggregation and our defense of it, it is necessary to first review the situationist critique more broadly. The situationist critique of global traits and virtues is premised upon two claims derived from empirical research in personality and social psychology (Fleeson et al. 2014; Jayawickreme et al. 2014). The claims are as follows:

1. *Situations exert apparently disproportionate influences on virtuous behavior.* The bulk of this dramatic-seeming evidence suggests that virtuous behavior is sensitive to small

contextual effects. Although the power of any single aspect of a situation is no larger than the effect of single traits (Funder and Ozer 1983), the cumulative effect of the multiple aspects of situations could be strong.

2. *Traits are weak predictors of behavior*. An often-weak empirical relationship has been noted between a person's behavior in one instance and their behavior in another instance, or between their standing on a measured trait and a single instance of behavior.

The situationist questions the viability of virtue ethics on the basis of these two premises. A review of the person-situation debate and the state of the relevant evidence is not our main focus, so our coverage of it is brief (please see Jayawickreme et al. 2014, Slingerland 2011, or Sreenivasan 2002, for a fuller review of the situationist argument, the evidence, and the defenses). If situations are so powerful, and traits so weak, the claim goes, character traits must be non-existent, highly localized and/or extremely weak. And if traits are nonexistent, highly localized, or weak then they are hardly a solid foundation upon which to build a viable ethical theory (Doris 2002; Harman 1999).

In this paper, we are defining traits as characteristics that individuals have levels of, usually or at least potentially differ on, result from an individual's specific history or biological makeup, are enduring over at least some time, and are relevant to an individual's actions, cognitions, emotions, or motivations (Fleeson et al. 2014). Traits have dimensional properties, such that people are ordered in the degree to which the trait content is descriptive of them. A person can be described very well by a trait content, moderately well, or not at all well, and in more fine increments of descriptiveness. For example, a person might be described by the extraversion content of talkativeness, assertiveness, and boldness very well, moderately well, or not well at all. In each case, the person has the trait at the specified level of descriptiveness.

Much empirical evidence presents some reason to believe that such traits exist and have meaningful effects on behavior (Jayawickreme and Fleeson 2017). First, scientific research has shown that traits are real: observers agree with each other about others' trait levels, traits exhibit impressive stability across adulthood, and there is evidence for a genetic component to traits (DeYoung 2010; Funder and Colvin 1997; Roberts and DelVecchio 2000). Second, traits matter: they are robust and consistent predictors of important outcomes, including academic achievement, job performance, marital quality, mental health, and even length of life (Ozer and Benet-Martinez 2006). Third, traits are broad: there is strong evidence that traits cover a broad range of behaviors, such that individuals who are higher than others on one type of behavior tend to be higher than others on a broad range of related behaviors. For example, individuals who are more assertive than others tend also to be more talkative than others, more adventurous than others, bolder than others, more active than others, and so on, leading to individual differences in the broad trait of extraversion. Fourth, there is now considerable consensus that traits are organized in a hierarchical structure, such as the Big Five traits of extraversion, agreeableness, conscientiousness, emotional stability, and intellect/open-mindedness (Soto and John 2017) or the HEXACO model (e.g., Ashton and Lee 2007), which adds a sixth trait of honesty/humility to the Big Five. These traits together describe many of the important individual differences in individuals' patterns of thinking, feeling, and behaving. It would be hard to come up with a consistent explanation other than the existence of broad traits to account for all of these findings.

Aggregated Behavior is Consistent One reason for the importance of the aggregation principle is that it provides the needed direct refutation of the situationists's inconsistency

evidence. Most researchers have accepted the empirical result that the correlation between any two behaviors performed by the same person, or between a measured personality trait and a single behavior, is often modest. Thus, the consistency needed for the defense of virtues does not seem to be present at the single action level. However, researchers did find strong evidence of consistent individual differences in behavior by aggregating across multiple assessments. According to the principle of aggregation, the sum or average of a set of multiple measurements is more stable and unbiased than any single measurement from that set. For example, Fishbein and Ajzen (1974) found that attitude scales correlated around .70 to .90 with behaviors that were aggregated. Similarly, Epstein (1979) argued that aggregated behavior was indeed highly consistent for a variety of personality traits, and his view was highly influential in promoting the importance of aggregation. Epstein was not working with traits explicitly identified as the Big Five, but his point was intended to be general to all traits.

Thus, aggregation is the key to establishing the relationship between traits and behavior, and demonstrating the meaningfulness of global traits. Indeed, the person-situation debate occurred, at least in part, because of a failure to recognize that traits refer most fundamentally to broad, stable, "week-in, week-out" consistent patterns of behavior across representative samples of situations (Epstein 1979). Thus, the nature and existence of traits should be evaluated in the context of aggregated behavioral tendencies rather than by examining relations between single items of behavior.

2 One Defense of Global Traits and Virtues: Whole Trait Theory

Whole Trait Theory (Fleeson et al. 2014), building on several previous approaches (e.g. Buss and Craik 1983; Epstein 1979; Shoda et al. 1994), provides a model of traits that incorporates aggregation into its defense of global traits and that directly addresses the question of variability in behavior. According to Whole Trait Theory, when traits are used to describe what people do, traits are describing individuals' entire distributions of the different ways each individual acts.

Whole Trait Theory starts with the concept of a "personality state" (Fleeson and Law 2015; Cattell et al. 1947). A state has the same content as a corresponding trait, but applies for a shorter duration. For example, an extraverted state has the same content as extraversion (talkativeness, energy, boldness, assertiveness, etc.), but applies as an accurate description for only a few minutes. States exist along the same dimensions as do traits, with the descriptiveness of the content varying in degrees from high to low. State content, like trait content, is primarily behavioral, but it also includes some cognitive and affective content. The personality state concept facilitates a comprehensive assessment of each individual's behavior in naturally occurring situations.

Across a short of period of time, a person will enact states at various points along the dimension. Sometimes he or she will enact a high level of the state and other times a low level of the state. All of these states together will form an individual distribution of states for that person. For example, over a week, a person might act highly agreeable at times, moderately agreeable at other times, and possibly disagreeable at still other times. The number of times the person acts at each level of agreeableness can be combined to form a distribution of agreeableness states for that person.

Whole Trait Theory does, it should be noted, support the observations of situationism. A typical individual varies in his or her behavior across two weeks about the same as the amount

a typical individual varies in his or her affect across two weeks, and more than the amount people differ from each other in their average levels of traits, meaning that individuals differ from themselves more than they differ from others. Thus, when observing the same people in multiple situations, and even when obtaining direct measures of the Big Five traits, a density distribution approach directly verifies the high levels of within-person variability that were indirectly implied by the results in earlier experiments.

While the evidence for Whole Trait Theory verifies some arguments proposed by situationists, it also strongly supports the trait perspective. That is, consistency does exist, but in the distribution as a whole and not in single states (Epstein 1979; Fleeson and Law 2015). Most importantly, although each individual's distribution is very wide, different people's distributions occupy *different locations on the dimension*. This location can be represented by the distribution's central point or tendency around which he or she varied. For example, although a person may act at different levels of agreeableness states across time and situations, his or her entire distribution of states might lie on the high end of the agreeableness dimension, on the low end, or right in the middle of the dimension. The central point of a person's distribution is a convenient way of summarizing where on the dimension the person's distribution lies.

The key question is whether each individual's central point remains stationary across time. If the central point remains stationary, then people differ in their level of a given virtue. And if the central point remains robustly stationary, then people differ in their level of the given virtue consistently. The evidence has revealed that people do differ in their central points and that they consistently have the same central point as themselves over time. For example, the person whose distribution is to the agreeable side and thus has a central point of agreeableness that is fairly high would have a central point of agreeableness that is fairly high week and after week after week. Indeed, differences between individuals in their location were as consistent and predictive as any variable in psychology.[1]

Thus, stable personality differences are robustly apparent in people's overall distributions, even if not in their single behaviors. What is important for this paper is how those central points are arrived at: via the principle of aggregation. The central point of a person's distribution of states is precisely the aggregated average way the person acted across the distribution. This aggregation is what provides the needed consistency of traits. Thus, aggregation appears to provide the consistency needed to establish virtues. This apparent consistency is the point we return to later.

Note that this view of traits built on but was forced to modify the standard view that many might have held prior to the situationist challenge to traits. Whole Trait Theory built on standard models by emphasizing the average way someone acts, but modified standard models by incorporating the important discoveries of the situationists – and this approach could be integrated into new theorizing about virtue (Jayawickreme and Fleeson 2017).

One assumption in virtue theory is that people differ from each other in at least some substantial, regular, and meaningful way in their behavior. That is, the validity of virtue theory depends upon the existence of at least one kind of consistency. The facts that a) people's distributions of behavior do indeed differ systematically from one another, b) these

[1] Degree of remaining stationary was assessed by calculating consistency of differences between people. In these data consistency was evident because central points remained in approximately the same location across time. Please see Fleeson and Furr (2016) for further discussion of why consistency of individual differences is essential to consistency of individuals.

distributions of behavior are impressively stable, c) these individual differences have substantial implications for the quality of one's life, including happiness, longevity, income, health and relationships, and d) conceptualizing traits as distributions of trait-relevant behavior successfully reconciles the trait perspective with evidence for substantial within-person variability, push the matter in favor of the existence of robust character traits.

Moreover, it is clear that while distributions are real descriptions of how people act, *they need to be produced by mechanisms capable of discriminating between situations*. Whole Trait Theory here provides an explanatory account of traits (Fleeson and Jayawickreme 2015; Jayawickreme and Fleeson 2017). The task of the explanatory account is to explain the distributions – that is, to explain why people differ from each other in their distributions (origin of traits) and to explain the within-person variability in states within the distributions (mechanisms constituting traits). The explanatory account of the Big Five distributions is needed to accomplish these two tasks. Adding an explanatory account to the Big Five creates two parts to traits, an explanatory part and a descriptive part, and these two parts are separate but also are joined into whole traits.

Whole Trait Theory proposes that this explanatory side of traits consists of "social-cognitive" mechanisms such as goals, plans and rules (Mischel 2004; Snow 2010). This is because social-cognitive mechanisms are clearly important in personality, and because density distributions of states make it clear that personality is responsive to situations. We propose that several such processes are the determinants of states (Fleeson and Jolley 2006). These processes include interpretative processes, motivational processes, stability-inducing processes, temporal processes, and random error processes. We note here, as we have elsewhere (Jayawickreme and Fleeson 2017), that having the appropriate motivation and beliefs are indeed important for possessing virtue. That being said, the situationist criticisms of global traits originally relied on evidence concerning consistency of *behavior* (e.g. Doris 1998). One could argue that consistency of behavior is an important prerequisite for even considering the question of whether virtue exists, since without consistent behavior, examining consistency of motives and beliefs would be moot.[2] By presenting evidence of consistency of behavior, we can then move to the next step of testing for the consistently of relevant motivations and beliefs.

3 Doris' Criticism of Aggregation: The Heinous, Disqualifying Action

We now turn to a significant worry about aggregation, a worry articulated by Doris. Doris (2002) displayed great thoroughness and perspicacity by anticipating the aggregation defense and responding to it. For this he deserves credit, as the principle of aggregation is not an easy one for psychologists to fully appreciate, let alone for individuals not primarily trained as psychologists. Indeed, his analogy to math tests casts aggregation in almost as favorable a light as possible. In this analogy, Doris correctly points out that we would never expect a single math test item to reveal an individual's math skill, because multiple factors determine performance on a single item. Only with multiple items could we adequately assess a person's

[2] This is admittedly not entirely true, as it is quite possible that practical wisdom could lead one to act apparently inconsistently at a purely behavioral level based on a consistent but complex application of virtues to diverse and complex situations. Such types of defenses of virtue will be needed if aggregation falls. However, we believe such a defense is not a position that needs to be taken because we contend that aggregation does not fall.

level of math skill. Likewise, we should not expect a single behavior to reveal someone's level of the corresponding virtue, because multiple factors determine performance of a single behavior.

Despite this clear depiction of aggregation, Doris unfortunately does not describe his objections to aggregation in an equally clear manner. We will therefore take the liberty of doing our best to faithfully represent the strongest argument he might be making.[3] The strongest argument, in our opinion, and the one that best represents this compelling concern, is also the most complex. In brief, this argument is that aggregation gives up on predicting single actions, but that single actions are the key objects needed to be predictable when it comes to virtue. The single actions that *cannot* be ruled out by aggregation-based traits are precisely the kinds of single actions that *are* ruled out by virtues. That is, it is possible that the people identified by aggregation as virtuous do enact heinous actions, but heinous actions disqualify a person as virtuous. Thus, aggregation is not a reliable identifier of virtuous individuals.

Doris points out that "All situations are not created equal, and people care more about some situations than others; the aggregationists' protest notwithstanding, people attend closely to particular situations of concern" (p. 75). The situations he is referring to are ones in which a person might commit a very bad action, or even a heinous one. For example, "what my faithful partner and best friend will be up to if I come home an hour early from work" (p. 74), or whether the babysitter will "molest our children next Tuesday night from seven to eleven when we go out for dinner and a show" (p. 74) or what one's lover "will do when faced with a particular instance of sexual temptation" (p. 74). Doris implies that even one such single heinous behavior is incompatible with the virtue. One single instance of cheating is enough to disqualify loyalty; one single instance of molesting disqualifies trustworthiness; one single act of murder disqualifies compassion.

However, the argument goes, because aggregates are only weakly correlated with any specific behavior, they cannot rule out the enactment of single heinous behaviors. Since aggregates are what had appeared to provide the consistency of behavior necessary for virtues, but aggregates here are shown not to provide this needed consistency, they can no longer serve that function and can no longer be the basis of virtues. Some other defense of virtues would be needed.

For example, imagine a person who has an aggregate level of compassion that is very high, but also committed one instance of unjustified murder. That one instance of murder would disqualify that person from having the virtue of compassion. No number of other acts of compassion would be capable of cancelling out the one murder. If aggregates cannot rule out that this one instance of murder might happen, they cannot give us much confidence in the compassion of the individual even though he or she has the aggregate high compassion..[4]

[3] Doris's objections appear to form three distinct arguments. The first is that aggregation is controversial and has significant methodological issues. The second argument is that the predictions made available by aggregation are only "tepid", whereas characterological moral psychology requires robust predictions. Please see Jayawickreme and Fleeson (2017) for a fuller account of these objections, our grouping of them into three arguments, and our responses explaining why aggregation is not controversial nor are the predictions tepid. However, in this paper, we focus on the third argument, because we believe it is the most worrisome. We also note that this and the next two sections of the paper build on and elaborate arguments we made in that chapter.

[4] This argument assumes that there is no change in character. Change in character would complicate the picture, because the heinous action might not be committed by the "same" person who enacted the compassionate actions. This is an interesting possibility for another paper.

The reason that aggregates are believed to allow such heinous behaviors is that aggregates can predict single behaviors only weakly. Empirical results suggest that the correlation might be around .20, such that a person's aggregate puts only a weak constraint on the person for any given single action. As a heinous action is a single action, aggregates put only a weak constraint on whether the person will commit a heinous action. Although aggregates might be good enough to justify non-moral traits, these arguments imply that aggregates are not sufficient to justify virtues. Doris succinctly concludes: "Crucially, Epstein's traits are not the ones in characterological moral psychology" (p. 74).

In sum:

a. Robust virtues are incompatible with heinous actions.
b. People designated as virtuous by aggregation might perform heinous actions.
c. Thus, the people designated as virtuous by aggregation cannot count as virtuous.

For the purposes of this paper, we grant the logic of the syllogism, and we grant premise (a).[5] Rather, much of the remainder of our response will focus on (b). The key premise in Doris' rejection of the aggregation solution, from our point of view, is this premise (b), that disqualifying, heinous events are committed by people with high aggregates. This assumption is what makes aggregation unsuitable as the undergirding for virtues.

If premise (b) is correct, then granted premise (a) conclusion (c) follows, that people designated as virtuous by aggregation cannot count as virtuous. Thus, aggregation cannot be the basis of defense of virtue theories against situationism, and since many defenses of virtue rely on aggregation, (e.g. Jayawickreme et al. 2014; Slingerland 2011), those defenses are undermined.[6] Since Doris' rejection of the aggregation solution rides on whether aggregates are compatible with heinous exceptions, much rides on whether aggregates are compatible with heinous exceptions.

4 The Defense of Aggregation

In responding to this concern, our first step is to divide single actions into two sorts. We believe that much of the power of this concern comes from conflating two kinds of single actions. The two kinds of single actions are mundane single actions and heinous single actions. To his credit, Doris makes this distinction, when he points out that "not all situations are created equal, and people care more about some situations than others" (p. 75). We assume that there are a range of single actions from heinous to morally commendable. On the immoral end, actions range from heinous to very bad to modestly bad. For example, consider the dimension of compassion. Compassionate acts can range from heinously uncompassionate to gloriously compassionate. Murder might be a heinously uncompassionate action, physical violence might be a very bad action, failing to volunteer for a needed task might be modestly uncompassionate, holding the door for someone might be modestly compassionate, working with habitat for

[5] However, we note that several writers have taken issue with some version of premise (a), including Slingerland's (2011) thoughtful discussion on how Confucian accounts of virtue ethics get beyond the strict "high bar" of absolute virtue demanded by Doris and the situationists.
[6] Although other defenses of global virtues and traits may still stand.

humanity might be very compassionate, and donating a kidney to an anonymous stranger might be gloriously compassionate.

When it comes to mundane single actions, we in fact agree with Doris that people do not care very much about them, and we also argue that virtues do not appear to be disqualified by mundane single actions that are not entirely expressive of virtues. The mundane actions, near the midpoint of the dimension, are not very important. We are not surprised if a very compassionate person doesn't volunteer for every possible action, or doesn't give all of his or her money away. The fact that aggregation allows such single exceptions does not undermine aggregation as a basis for virtue.

The heinous actions, rather, are the ones that give us pause. Babysitters molesting children and acts of murder are the sorts of rare and extremely bad actions that would make us question the virtue of the actor. These are the sorts of actions that would seem to undermine aggregation as a reasonable defense – if aggregation cannot rule out such actions, then aggregation surely does not provide sufficient basis for virtue.

4.1 Heinous Exceptions do Not Happen – Case I

The critical question is therefore: Is it right that aggregates are nearly useless in predicting whether someone will commit a heinous or very bad action? The main claim of our paper is that they are not. The reason this objection to aggregation fails is that aggregation is actually not quite so open as it seems it should be. That is, aggregates do not allow all actions equally, and in particular, are not so open to actions that are at the other extreme end of the virtue dimension.

Our argument is that people identified as virtuous by aggregates alone will almost never commit heinous actions. The reason for our argument is that correlations do indeed take into account the extremity of the deviations. In other words, correlations do not treat all single behaviors alike. A correlation of .20, while relatively small in many cases, is actually quite strong in predicting the non-occurrence of such extreme deviations. The general version of this point and the underlying mathematics were first introduced by Taylor and Russell (1939).

To show this concretely, we used a simulation. A simulation can be described as a large scale thought experiment, in which examples are created to take specified assumptions and follow them through to their consequences. Simulations do however differ from thought experiments in that many examples are worked through at once, the thought experiment is conducted by a program, and a simulation can easily encode assumptions of unrelated determinants by designating them as randomly determined. The simulation encodes only the information contained in the assumptions, and nothing else. It translates the assumptions into precise numbers and distributions. Both this encoding of assumptions and translation of assumptions into numbers are open so their accuracy can be checked by readers. A simulation allows other and unspecified determinants of actions because it can set actions to be randomly distributed, which represents other causal forces that are unrelated to the encoded forces. Thus, a simulation consists of setting up the assumed parts, letting there be unspecified parts by allowing them to be random, and then calculating the consequences of the assumptions.

In the simulation, we created aggregate levels of compassion for 10,000 simulated people. The levels were created randomly, with the only information about each individual being their distinct, randomly generated aggregated level of compassion on a compassion dimension. Table 1 lists the mechanics of the simulation, the corresponding assumptions, and the justification for the assumptions.

Table 1 Clarification of the simulation and its connection to theoretical assumptions and justifications

Mechanics of simulation	Corresponding assumption	Justification
1. Created one piece of information about each simulated person, which is the person's aggregate level of compassion.	1a. People differ in aggregate levels of compassion. 1b. Aggregates represent the person's average level of compassion. 1c. A person's aggregate level of compassion can be represented as a number on a compassion scale. 1d. **The aggregate level is enough information to confidently predict whether the person will commit a heinous action.**	1a. Empirical results accepted by Doris. 1b. Empirical results accepted by Doris. 1c. Compassion is dimensional. 1d. **This is the key assumption of the aggregation defense of virtues, and is the key assumption being tested by the simulation.**
2. Randomly created normal distribution of aggregate levels, one for each person. 2a. set average to 0 and standard deviation to 1.	2. Different people have different levels of aggregate compassion and these levels are distributed normally.	2. Empirical results accepted by Doris. Personality and behavior differences are typically found to approximate the shape of normal distributions. 2a. Mean and standard deviation are arbitrary and inconsequential, so we picked easy numbers.
3. Created 10,000 simulated people.		3. A large number but still manageable.
4. Grouped people into moral categories based on aggregate level.		4. Cutoffs represent discussed moral categories, and are intended to ease presentation, but precise cutoff points do not affect results.
5. Create a single action for each person. 5a. distributed normally with average set to 0 and standard deviation to 1. 5b. correlated .20 with aggregate level. 5c. otherwise randomly determined.	5. People conduct single actions that differ in degree of compassion on the same scale. 5b. **The only connection of a single action to the aggregate is the .2 correlation.** 5c. The remaining determinants of an action's degree of compassion are unrelated to the person's aggregate level, and such determinants include factors like the situation.	5a. Behaviors are typically found to approximate the shape of normal distributions. 5b. **This is the basis of the key argument against aggregation.** 5c. Random determination of the remaining degree of compassion of the actions represents the actions being determined by forces other than and unrelated to the person's aggregate level. This is the corollary of the key argument against aggregation.
6. Created 25,000 such actions for each person.		6. 250,000,000 total actions seemed sufficient.
7. Grouped actions into moral categories by degree of compassion.		7. Used multiple categories to allow reader preference to guide interpretation.

The first 5 steps were for the first simulation, and the final two steps were for the second simulation

On our compassion dimension, actions with positive numbers were considered compassionate and the actions with negative numbers were considered uncompassionate. The more positive, the more compassionate, and the more negative the number, the less compassionate. The simulated individuals' aggregate compassion scores ranged from −3.68 to 3.71. Half of them had an aggregate compassion level of 0.01 or above. 15% of them had a compassion aggregate score of 1.03 or above – we might tentatively call these the virtuous. The top 2% had an aggregate compassion score of 2.00 or above. We might tentatively call these the morally exceptional.

For each of these people, we then generated a single action of varying compassion level. The only restriction on those single acts was that they be correlated about .20 with the average degrees of compassion by the same person and that they were normally distributed. This is the weak relationship between aggregation and single behaviors that raises so many questions about the strength of virtues in determining behavior. The other 96% of the determinants of the behavior we left to other factors, such as the situation. The simulation represents these other factors as resulting in a normally and randomly distributed set of actions. Doing so represents all other determinants of actions in the simulation and guarantees that the actions are otherwise unrelated to the aggregates. It does not imply that the actions are randomly determined, but rather implies that they are determined by forces that are unrelated to the aggregate and that result in roughly normal distributions of behavior. Thus, setting the other determinants to result in a normal and random distribution of actions makes the case against aggregation the strongest. Essentially, some simulated people were faced with very easy situations in which to act compassionately and others were faced with very trying situations, but the situations they faced were not related to their aggregate level of compassion.

In creating these actions, we assumed the normal, bell-shaped distribution of actions. A normal distribution means that more extremely bad actions are less frequent than less extremely bad actions. For example, murder is less frequent than attempted murder, which is less frequent than assault, which is less frequent than screaming, which is less frequent than dirty looks, and so on. It is possible that a society could exist in which the more extreme acts are equally or more frequent than the less extreme acts. But in such cases, the .2 correlations between actions would likely be much higher, consistency would not have been questioned, and the situationist challenge would not have occurred in the first place. That is, we assumed the normal distribution to approximate the conditions that led to the original situationist challenge. The single compassionate actions ranged in value from −4.18 to +3.67. The morally worst actions were those with the greatest negative numbers and the morally best actions were those with the greatest positive numbers.

We did this multiple times, and selected a typical result that was fairly common. We present the results in a scatterplot in Fig. 1, with the aggregated average level of compassion along the bottom and the degree of compassion in the single act along the vertical left side.

Because the y-axis along the left side shows the degree of compassion, the acts very low in compassion are at the bottom of the figure. Probably murder would be even lower on the scale than is depicted in this figure, but for the purpose of this example, we will allow murder to be at the very bottom of the scale. The people very high in average, aggregated compassion are on the right side of the figure. Each dot in the figure represents one person, with a combination of a certain degree of average compassion and a certain degree of compassion in the given single action.

Putting this together creates a "region of possible disqualification," in which people with very high degrees of average compassion nonetheless enact a very non-compassionate behavior. Such behaviors might be so heinous that they disqualify the person from being counted as compassionate, even though they had acted in a very compassionate way in many other instances.

The important point to see here is that there are nearly no people in the region of disqualification (exactly none in this particular example). That is, even though the correlation is "only" .20, such a correlation is robust enough to have confidence that the highly compassionate person – based on their aggregate compassion – will not engage in heinous, disqualifying acts.

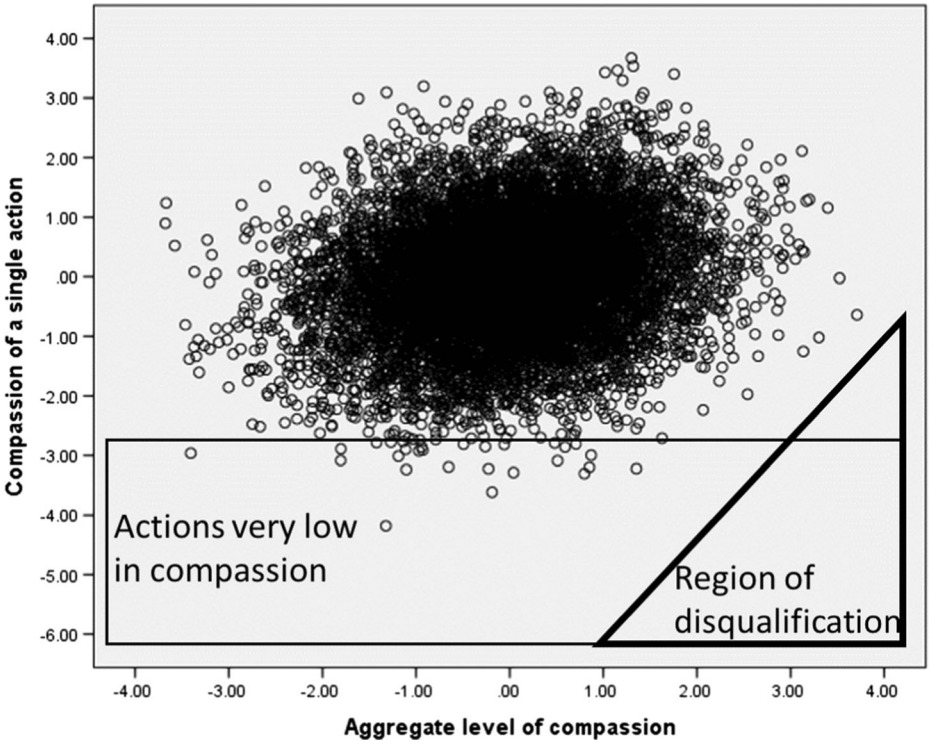

Fig. 1 Aggregates do strongly predict a lack of disqualifying actions

The prediction does not mean that it will never happen. We cannot have certainty that none of the highly compassionate people will not have a breakdown and commit murder. Life is indeed somewhat uncertain. But the observed correlation means that is extraordinarily rare. We can count on people with high degrees of aggregated virtue to not enact any disqualifying behaviors. These are indeed the robust traits moral psychology requires.

4.2 Heinous Exceptions do Not Happen – Case II

Doris may object that our analysis is misguided in one regard. Our analysis shows the single actions of many individuals, but Doris' rejection of aggregation concerns the many actions of one individual.[7] To respond to this objection, we need to generate many actions for each individual. We then take the most morally exceptional individuals, according to aggregation, and see how many of these morally exceptional individuals committed heinous actions. We also see how many of them committed very morally bad actions.

This test starts with the same people as in the first example above, but instead of one action per person, creates many actions per person. How many actions per person are necessary? How many heinous-free actions must a person perform in order to not disqualify his or her standing as virtuous? 1000 actions is a lot of actions, and might be enough. Perhaps 10,000 actions would be enough to see whether a virtuous person would commit a heinous action. We

[7] We thank Ryan West for this point.

decided to allow each person 25,000 actions – that is, 25,000 chances to commit a heinous action. This resulted in a total of 250 million actions, of varying degrees of compassion.

When generating the 25,000 actions for a given individual, the only constraints we put on the degree of compassion of the individuals' actions were the individual's aggregate level of the virtue and the "weak" .20 correlation between said aggregate level and the behavior and a normal distribution, as before. Because the other 96% was random, representing the situation effect, and because each individual performed 25,000 actions, individuals differed quite a bit in the situations they faced, just as in real life. Some of the 10,000 simulated individuals faced some very difficult situations, and some of them even experienced a string of very difficult situations.

Which actions count as heinous? The worst actions are the actions that have the most negative numbers. Of the 250 million actions, the least compassionate action was −5.96, and actions became increasingly compassionate from there, becoming less and less negative, through neutral, to having positive scores on compassion. Out of all the actions people perform in their daily life – from brushing their teeth to driving to work to removing a stick from the path to saying hello to a colleague to setting up their computer to getting some work done to talking in a meeting, and so on, the kind of heinous actions mentioned by Doris – heinous actions such as molesting children or murder – are relatively rare among them.

We started by considering only the very worst actions as the heinous (less bad actions will be considered below). Out of 250 million actions, perhaps only 20 or so would reach the disqualifying level of badness. This number may actually be too high – we wouldn't expect 20 people out of 10,000 to commit murders[8] – but we will be generous to Doris' objection in our decisions. Thus, we included the worst 21 actions as the heinous actions in this sample. These were actions that had compassion levels of −5.26 or worse.

However, to avoid resting the argument on one admittedly arbitrary designation of bad actions, we used several other cutoffs for disqualifying actions. We give labels to the actions selected by these cutoffs in order to facilitate memory. However, we emphasize that these labels are defined solely by decreasing levels of badness. Each label consists of less bad actions than the previous. We do this to allow each reader to determine his or her own cutoff for disqualifying actions. The important point to consider is how bad an action has to be to disqualify the person committing the act as virtuous.

We took the slightly less bad actions as the "nearly heinous". These were the 50 actions with a compassion level between −5.00 and −5.25. We also considered "outrageous", "very morally bad actions", "bad actions", and "untoward actions" as those that were successively less bad. Table 2 shows the cutoffs and frequencies of each of those categories. Note that the numbers of actions in a given category increases rapidly. As the cutoff loosens, we quickly move away from the kind of horrible and very rare action that Doris was concerned about, and are documenting the bad kinds of actions that people enact on a more regular basis.

We can now turn to determining who the people are who commit the heinous and outrageous actions. Is it as likely to be the people with virtuous aggregates as those without virtuous aggregates, as Doris' interpretation of aggregation would expect? Or is it those with the lower aggregate virtue scores, as the defenders of aggregation would expect?

To find out, we took the people who are virtuous and the people who are morally exceptional according to the aggregates, and counted the number of heinous, nearly heinous,

[8] For example, the FBI reports that 4.5 per 100,000 people committed murder in 2014, a rate 1/40th of what we are allowing here (FBI, n.d.)

Table 2 Bad actions committed by good people

Type of Action	Number	Compassion score	Number of People Performing Action			
			Morally Exceptional $N = 250$	Fully Virtuous $N = 1307$	Morally Good $N = 3531$	Morally Indifferent $N = 4957$
Heinous	21	< −5.26	0	0	3	18
Nearly heinous	50	−5.25 < −5.00	0	2	3	44
Outrageous	759	−4.99 < −4.51	1	19	105	536
Very morally bad	7207	−4.50 < −4.00	10	208	1020	3056
Bad[a]	331,505	−3.99 < −3.00	4	616	3368	4957
Untoward[b]	5,355,589	−2.99 < −2.00	0	602	3531	4957

Table entries are numbers of people who performed at least one of the given type of action. The numbers may not add up if an individual performed the given action type more than once. Because standard deviations of the 25,000 action distributions were almost exactly 1.00, a given act would be a number of standard deviations below the mean equal to its number (e.g., −5.26 is about 5.26 standard deviations below the mean)

[a] Because nearly everyone performed at least one bad action, this row lists the number of people who performed more than 10 bad actions.

[b] Because nearly everyone performed several untoward actions, this row lists the number of people who performed more than 250 untoward actions.

outrageous, very morally bad, bad, and untoward actions that these people did. The first row of Table 2 shows who performed each of the heinous actions. Not a single one of those heinous actions was performed by a morally exceptional individual, even though each morally exceptional individual enacted 25,000 actions.[9] Furthermore, none of the virtuous individuals committed a heinous action either. Only 3 of the 3531 morally good individuals committed a heinous action, and nearly all (18 of 21) heinous actions were enacted by morally indifferent individuals. These 21 heinous actions make the logic of this argument very clear. When heinous actions are traced back to the actors, it is not to those with high aggregate levels of the underlying virtue; rather, it is to those with low aggregate levels of the underlying virtues. Thus, even though aggregates are not good at predicting single, mundane actions, they are quite good at predicting the lack of single, heinous actions.

We continued this logic with the nearly heinous, outrageous, very morally bad, morally bad, and untoward actions. The second row of Table 2 shows that not one of the "nearly heinous actions" was committed by the morally exceptional either. Here we see that two of the 1307 virtuous individuals enacted one nearly heinous action. Rather, the nearly heinous actions were performed by morally good (three people) and morally indifferent (44 people). When we get to the outrageous actions, here we see one morally exceptional individual fall. Only 19 of the 1307 virtuous individuals enacted an outrageous act, and only 105 of the 3531 morally good individuals performed even one outrageous action.

As we continue through successively less bad actions, the same pattern holds. For bad and untoward actions, we had to divide people by whether they committed groups of such actions, as opposed to single such actions, and the results were equally illuminating.

These findings make it very clear that aggregates do indeed rule out heinous and even the not-quite-heinous actions. Although the correlation appears to be weak, aggregates are quite

[9] We ran these simulations many, many times to verify their robustness. In one of those many runs, two of the heinous actions were performed by the morally exceptional individuals.

good at predicting the extreme actions – and the extreme actions are the important ones. Heinous actions are in fact not compatible with aggregates.

That said, we do admit that occasional exceptions did happen. A very small number of the virtuous individuals committed a nearly heinous or outrageous action amongst their 25,000 actions. Thus, a highly virtuous aggregate is not an iron-clad guarantee that the person will never commit a very bad action. However, only a very few such virtuous people did commit such an action. Indeed, we have come a long way from the apparently "weak" single-action correlations, to the point that we are somewhere just short of "ironclad" in the strength of our predictions.

We also repeated this simulation, but allowed a correlation of .35 between aggregates and single actions. A .35 correlation is a possibly more realistic account of the true correlation than is .2, given many empirical findings (Fleeson and Gallagher 2009; Funder 2001). In this case, the findings were even starker. For example only 36 of the morally exceptional individuals conducted even a bad action.

We acknowledge that these are only simulated data. However, simulated data have at least two advantages that are essential for an investigation such as this. First, simulated data allow us to specify that only the aggregate score and the correlation are responsible for the individual actions (e.g., there are no confounding factors). Thus, we can be sure that it is a pure test of whether virtuous aggregates rule out heinous actions. The second advantage of simulated data is that it is relatively easy for anyone to copy our procedures and test our conclusions for themselves.

As a final note, we want to mention that these analyses probably overestimate the possibility of heinous exceptions, and thereby underestimate the power of virtuous traits. This is because these simulated data do not include any patterned responses to situations that individuals may have. As Snow (2010) points out, patterned responses are very likely an important part of virtues. A patterned response is a coherent set of behavioral reactions to a set of situations. The simulation assumed that behavior was entirely random except for the aggregate, which gave maximum chance for heinous actions. Pattern responses would reduce the randomness and thus reduce the chance for heinous actions. Patterned responses to situations would make the power of personality even stronger in its prediction of actions.

5 What Exactly Does Aggregation Show?

How does it make sense that aggregates have relatively low correlations with single actions, yet powerfully rule out extreme single actions? It makes sense because of individuals' density distributions of actions. What aggregates do is reveal the central point of a person's distribution of actions along the virtue dimension. Actions vary within people, because people are complex and respond to situations in reasonable ways. Thus, distributions cover wide ranges of the scales. But distributions also differ between people. Distributions of people with high aggregate scores are located on the high end of the virtue scales.

These distributions are similar to the shapes of normal curves. The closer an action is to the central point of a distribution, the more frequently the person performs that action. The farther an action is from the central point of the distribution, the less frequently the person performs the action. Thus, there will be fewer and fewer of the actions that are farther and farther from a

person's central point, to the point of vanishing. If a person's central point is on the virtuous end, then the less virtuous the action, the fewer instances in which the person will perform the action, to the point of vanishing.

To return to a point touched on earlier, an additional feature of the wide distributions characteristic of people's actions is that they imply that the causes of the behaviors are motives and beliefs. Within-person variability is most likely explained by social-cognitive mechanisms such as motives and goals (Fleeson and Jayawickreme 2015; Snow 2010). In other words, aggregation is the best solution because it actually begs for analysis of motivations and beliefs to explain the causes of the within-person variability present in the aggregated distribution. In other words, the necessity to defend global virtues has led to the reconceptualization of global virtues as traits that makes them more amenable to such analyses. Such a move echoes a claim that psychologist Gordon Allport made almost 50 years ago:

> "To the situationist I concede that our theory of traits cannot be so simpleminded as it once was. We are now challenged to untangle the complex web of tendencies that constitute a person, however contradictory they may seem to be when activated differentially in various situations." (Allport 1968, p.47)

These "tendencies" that Allport pointed to were energized by the appropriate motivations and beliefs that are an important part of possessing virtue. Such an exciting project should form the basis for future research on moral personality (Jayawickreme and Fleeson 2017) and we can thank Doris and the other situationists for providing the stimulus for engaging this new and exciting direction.

6 Conclusion

One of the principles employed to claim that global traits are important is the principle of aggregation. Without this principle, the empirical evidence for behavioral variability seems to strongly undermine the impact of virtue and character on behavior. Doris laid out a formidable attack on the adequacy of aggregation. As there has been no detailed attempt to respond to this formidable attack on the principle of aggregation, aggregation as a defense of broad virtues has been doubtful. As we have shown through the use of simulated data, Doris' criticisms regarding aggregation do not stand up to scrutiny, and his arguments are ultimately unfounded.

However, we believe that it is to Doris' credit that he addressed the question of aggregation in the first place. The questions that he tackled are difficult and complex, and require the unpacking of complex statistical details. In doing so, we acknowledge Doris' concern that people who possess the virtues as understood by Aristotelian virtue ethics meet a "high bar" threshold when it comes to disqualifying actions. In our response to his attack, we took this this claim seriously. When it comes to the "high bar" of heinous exceptions, we showed that the aggregation defense of global virtues is in fact not undermined. Even though the correlation between aggregate averages and single behaviors might be as low as .20, aggregated averages do not admit the disqualifying actions. We believe, thus, that defenses of global traits that rely on aggregation are not in fact weakened. Nevertheless, we are thankful to Doris for engaging the question of aggregation in the first place, and pushing for a defense of global virtues

that necessitates their reconceptualization as traits that allow for explanation of both stability and variability in behavior.

Acknowledgements We thank Mike Furr, Ryan West, and Anna Hartley for comments on this line of reasoning. This paper was made possible through the support of a grant from the Templeton Religion Trust. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the Templeton Religion Trust.

References

Alfano M (2013) Character as moral fiction. Cambridge University Press, Cambridge

Allport GW (1968) The person in psychology: selected essays. Beacon Press, Boston

Ashton MC, Lee K (2007) Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. Personal Soc Psychol Rev 11:150–166

Buss D, Craik K (1983) The act frequency approach to personality. Psychol Rev 90:105–126

Cattell R, Cattell A, Rhymer RM (1947) P-technique demonstrated in determining psycho-physiological source traits in a normal individual. Psychometrika 12:267–288

DeYoung CG (2010) Personality neuroscience and the biology of traits. Soc Personal Psychol Compass 4(12): 1165–1180. doi:10.1111/j.1751-9004.2010.00327.x

Doris J (1998) Persons, situations and virtue ethics. Noûs 32:504–530

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

Epstein S (1979) The stability of behavior: I. On predicting most of the people much of the time. J Pers Soc Psychol 37(7):1097–1126. doi:10.1037/0022-3514.37.7.1097

Federal Bureau of Investigation (n.d.) Retrieved June 10, 2016, from https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/tables/table-16/Table_16_Rate_Number_of_Crimes_per_100000_Inhabitants_by_Population_Group_2014.xls

Fishbein M, Ajzen I (1974) Attitude towards objects as predictors of single and multiple behavioral criteria. Psychol Rev 81:59–74

Fleeson W, Jolley S (2006) A proposed theory of the adult development of Intraindividual variability in trait-manifesting behavior. In: Mroczek DK, Little TD (eds) Handbook of personality development. Lawrence Erlbaum Associates Publishers, Mahwah, pp 41–59

Fleeson W, Gallagher P (2009) The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. J Pers Soc Psychol 97(6):1097–1114

Fleeson W, Jayawickreme E (2015) Whole trait theory. J Res Pers 56:82–92. doi:10.1016/j.jrp.2014.10.009

Fleeson W, Law MK (2015) Trait enactments as density distributions: the role of actors, situations, and observers in explaining stability and variability. J Pers Soc Psychol 109(6):1090–1104. doi:10.1037/a0039517

Fleeson W, Furr RM (2016) Do broad character traits exist? Repeated assessments of individuals, not group summaries from classic experiments, provide the relevant evidence. In: Fileva I (ed) Questions of character. Oxford University Press, New York, p 231–248

Fleeson W, Furr RM, Jayawickreme E, Meindl P, Helzer EG (2014) Character: the prospects for a personality-based perspective on morality. Soc Personal Psychol Compass 8(4):178–191. doi:10.1111/spc3.12094

Funder DC, Ozer DJ (1983) Behavior as a function of the situation. J Pers Soc Psychol 44(1):107–112. doi:10.1037/0022-3514.44.1.107

Funder D, Colvin CR (1997) Congruence of others' and self-judgments of personality. In: Hogan R, Johnson JA, Briggs SR (eds) Handbook of personality psychology. Academic Press, San Diego, pp 617–647

Funder DC (2001) Personality. Annu Rev Psychol 52:197–221. doi:10.1146/annurev.psych.52.1.197

Harman G (1999) Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error. *Proc Aristot Soc* 99:315–332

Jayawickreme E, Fleeson W (2017) Does whole trait theory work for the virtues? In: Sinnott-Armstrong W, Miller CB (eds) Moral psychology, Volume 5, Virtue and happiness. MIT Press, Cambridge, pp 75–103

Jayawickreme E, Meindl P, Helzer EG, Furr RM, Fleeson W (2014) Virtuous states and virtuous traits: How the empirical evidence regarding the existence of broad traits saves virtue ethics from the situationist critique. Theory Res Educ 12:283–308

Mischel W (2004) Toward an integrative science of the person. Annu Rev Psychol 55:1–22. doi:10.1146/annurev.psych.55.042902.130709

Ozer D, Benet-Martinez V (2006) Personality and the prediction of consequential outcomes. Annu Rev Psychol 57:401–421

Roberts BW, DelVecchio WF (2000) The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. Psychol Bull 126:3–25

Shoda Y, Mischel W, Wright JC (1994) Intraindividual stability in the organization of and patterning of behavior: incorporating psychological situations into the idiographic analysis of personality. J Pers Soc Psychol 67:674–687

Slingerland, E. (2011). The Situationist critique and early Confucian virtue ethics. *Ethics*, 121(2), 390–419.

Snow NE (2010) Virtue as social intelligence: an empirically grounded theory. Taylor & Francis, New York

Soto CJ, John OP (2017) The next big Five inventory (BFI-2): developing and assessing a hierarchical model with 15 facets to enhance bandwidth, Fidelity, and predictive power. J Pers Soc Psychol. doi:10.1037/pspp0000096

Sreenivasan G (2002) Errors about errors: virtue theory and trait attribution. Mind: A Quarterly Review of Philosophy 111(441):47–68

Taylor HC, Russell JT (1939) The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. J Appl Psychol 23(5):565–578. doi:10.1037/h0057079

Vranas P (2009) Against moral character evaluations: the undetectability of virtue and vice. J Ethics 13:233–233