Can virtue be measured?

Randall Curren
University of Rochester, USA; University of Birmingham, UK

Ben Kotzee
University of Birmingham, UK

Abstract
This article explores some general considerations bearing on the question of whether virtue can be measured. What is moral virtue? What are measurement and evaluation, and what do they presuppose about the nature of what is measured or evaluated? What are the prospective contexts of, and purposes for, measuring or evaluating virtue, and how would these shape the legitimacy, methods, and likely success of measurement and evaluation? We contrast the realist presuppositions of virtue and measurement of virtue with the behavioral operationalism of a common conception of measurement in psychometrics. We suggest a realist and non-reductive conceptualization of the measurability of virtue. We then discuss three possible educational contexts in which the measurement of virtue might be pursued: high-stakes testing and accountability schemes, the evaluation of programs in character education, and routine student evaluation. We argue that high-stakes testing of virtue would be ill-advised and counterproductive. We make some suggestions for how program evaluation in character education might proceed, and offer some examples of evaluation of student virtue-related learning. We conclude that virtue acquisition might be measured in a population of students accurately enough for program evaluation while also arguing that student and program evaluation do not require comprehensive evaluations of how virtuous individual students are. Routine student evaluation will typically focus on specific aspects of virtue acquisition, and program evaluations can measure the aggregate progress of virtue acquisition in all its aspects while evaluating only limited aspects of the learning of individual students.

Keywords
Measurement, moral education, operationalism, program evaluation, student evaluation, virtues

Introduction
Can virtue be measured? Developments in moral education were long dominated by Lawrence Kohlberg’s cognitive developmental theory, and the success of moral education has accordingly been judged by its effect on developmental level or progress toward
reasoning from universal moral principles. James Rest’s (1979) Defining Issues Test (DIT) became a widely used successor to Kohlberg’s Moral Judgment Interview (MJI; Colby and Kohlberg, 1987), establishing itself as a quantified measure of level or quality of moral judgment. Are such tests indeed measurements of moral psychological attributes (Borsboom, 2005; Finkelstein, 2005; Maul, 2013; Michell, 1999, 2005, 2008)? Are the psychological attributes they claim to measure real (cf. Doris, 2002; Harman, 1999; Vranus, 2009), and, if real, are they understood as explanatory of patterns of behavior or operationally defined as the very patterns of behavior that moral attributes are commonly understood to explain (Borsboom, 2005; McDonald, 2013; Maul, 2013; Norris et al., 2004)? Does measurement require that these attributes have an inherently additive structure that justifies estimation of ratios, as measurement in the physical sciences requires (Michell, 1999, 2005, 2008)? If the answer is ‘yes’ and moral virtue is measurable, then it could make sense to say that a person grew 10% taller or more virtuous in the course of a year, or is twice as tall or virtuous as someone else. How virtue could be measurable in this sense is hard to imagine. Or are ordinal relationships of ‘more’ or ‘less’ of an attribute sufficient (McDonald, 2013)? If the DIT and related tests do indeed measure real psychological attributes, is it levels of moral judgment they actually measure? Supposing tests such as the DIT do indeed measure levels of moral judgment, is this sufficient for the purposes of moral education? If moral judgment, motivation, and conduct are not strongly linked, would the evaluation of moral education programs not require more than tests of moral judgment (Emler, 1996)?

It is this final question that is pointedly raised by the character education movement and its focus on virtue. If virtue or good character is the goal of moral education, and good character involves the possession of an integrated cluster of dispositions of desire, emotion, perception, belief, reasoned judgment, and action, moral educators will need measures of virtue beyond just measures of good moral judgment. In addition, they will need measures of moral sensitivity and moral motivation (at least), if not a measure of how moral sensitivity, judgment, and motivation all operate together to produce moral action. Hence, the question, ‘Can virtue be measured?’ To answer this, we need to consider what virtue is, what measurement is, and what kinds of instruments would be required to measure virtue.

It is no less important to ask what tools of evaluation moral educators actually require in order to do their work responsibly. Judging or evaluating people’s virtues or moral qualities is evidently something we all do and can scarcely avoid doing, if ‘evaluating’ means judging persons and their character ‘against an independent and objective standard of excellence’ (Wolff, 2007: 460). Evaluating students’ learning is a common, if not essential (p. 462), aspect of teaching, and the evaluation of students’ character may be important to some aspects of the administration of schools and student discipline. Whether such forms of student evaluation can or should constitute measurement is not obvious. Nor is it obvious that the evaluation of programs of virtue-focused moral education requires the measurement of virtue as such, as we discuss below. Evaluation of such programs might measure the distribution in a population of an attribute that is not possessed in measurable degrees by individual members of the population. If being virtuous were like being female, in being a categorical rather than quantitative attribute, the impact of educational and policy interventions on the distribution of virtue and underrepresentation of females in a
society might nevertheless both be measureable, the traits being additive at the level of a population (cf., Borsboom, 2005; Michell, 1999). Apart from the evaluation of educational interventions, the availability of serviceable measures of virtue would be essential to undertaking a variety of potentially valuable research studies.

Our aim in what follows is to make some progress in answering the various questions bearing on the measurability of virtue, by addressing the nature of virtue, the nature of measurement, and the purposes one might have in seeking to measure virtue. The purposes one has in mind in assessing virtue will substantially determine the choice of methods, how much success can be expected, and whether it is best to proceed at all. We will suggest some forms of student evaluation suitable to virtue-focused moral education, and suggest a mixed measures approach to program evaluation. Our brief foray into measurement theory will contrast the non-reductive realist presuppositions of virtue and measurement of virtue with the reductive behavioral operationalism that remains common in psychometrics, and we will suggest a realist conceptualization of the evaluation and measurability of virtue. We see no reason in principle why an individual’s character could not in some circumstances be evaluated, and in some sense measured, nor why virtue acquisition could not be adequately measured in the aggregate in a population of students for the purposes of program evaluation. We also argue, however, that the evaluation of student learning and programs in virtue-focused moral education do not require comprehensive measurements of how virtuous individual students are. In routine evaluation of student learning, one would ordinarily evaluate only limited forms of learning that may be indicative of progress in virtue acquisition, but without evaluating or measuring gradations of virtue acquisition as such. In evaluating character education programs, it would be most useful to adopt methods that aggregate scores from a combination of measures targeting different aspects of virtue in different samples drawn from the relevant student population. Such methods could, in principle, yield measures of the efficacy of virtue-focused education (in the strict sense supporting estimation of ratios; Michell 1999, 2005), without measuring (in any sense) how virtuous individual students are, since there would be no need to evaluate or measure more than one or two aspects of the virtues of individual students, or to compile aggregate scores for individual students even if two or more aspect scores were obtained for them. We also argue that the measures could rely on test item formats that are not machine-scoreable, and on test forms that are designed to yield scores with system-level significance but do not support summary or comparative judgments of individual students.

In asking whether virtue can be measured, we have in mind moral virtue generally and the various moral virtues that might be distinguished. Other desirable personal attributes, such as language mastery, expertise, speed, strength, and endurance, are measured routinely and without controversy. Is it conceivable that moral virtues might also be measured routinely and without controversy? We begin by considering the nature of moral virtues.

**Moral virtues**

The similarities between moral virtues and other forms of goodness are evident not only in accounts of moral virtue that emphasize their similarity to complex skills (Annas,
2011; Snow, 2010), but in what we know of the history of ideas of virtue in Greek antiquity. There are wider and narrower senses of the word ‘virtue’ and its ancient Greek counterpart, arête (signifying virtue, goodness, or excellence). In its wider sense, we speak of the virtues or good features of all sorts of things: the qualities that suit them for some purpose or make them pleasing or admirable. One virtue of a good hammer is that its striking surface is hard but not brittle, large enough to be a sufficiently wide target when swung toward a nail but not so wide as to visually obstruct a clear view of the nail, and placed directly in front of the center of gravity of the hammer head, so as to transmit the momentum of the hammer head’s mass to the nail efficiently. Attributes of persons can be virtues in this sense relative to specific activities or roles. Endurance is a virtue, or desirable attribute, in a runner of marathons, courage a virtue of a soldier, or a man, if soldiering is something required of men as such. The idea of human virtue or the virtues of a human being as such seems to have originated in this way in Greek antiquity, as the manly virtues associated with the defense of vulnerable polises or politically autonomous cities – strength, courage, cunning, and endurance – and later broadened to include attributes less obviously essential to the possessor’s success, but nevertheless essential to the internal functioning and stability of established societies, such as moderation, justice, and wisdom (Curren, 1996). Wisdom was understood to entail a kind of respect for reason and beings who reason, and thereby an ethic of mutual goodwill and norms of truthful reason giving. It was also understood, in one way or another, to be an essential component of being fully virtuous: true virtues, in the Aristotelian version of this idea, are guided by wisdom or good practical judgment.

An Aristotelian understanding of moral virtues views them as complex dispositions or clusters of related dispositions that are formed through habituation, which typically consists of both passive immersion in a good social environment and guided practice in acting well and becoming better (Curren, in press; cf. Steutel and Spiecker, 2004). Such habituation is understood to shape dispositions of desire, emotion, perception, belief, conduct, and reason responsiveness as a causally related package. It is hard to see how it could succeed without supervision and coaching that provides learners with a vocabulary of the good, calls their attention to factors that make a difference to how one should act, and guides them in exercising the forms of discernment, imagination, reasoning, and judgment on which good decisions are based. The point of practice is not simply for the learner to become reliably respectful of others and committed to good ends, but to develop the perceptiveness, imagination, judgment, and fortitude needed to actually achieve good ends. If it is true that moral virtues are dispositional clusters formed in this way, and moral perceptiveness and judgment are among the trailing effects of habituation that also shapes moral motivation and commitment, then evidence of ethical perceptiveness and judgment would be evidence that moral motivation and dispositions to act well are also present in those who have been morally educated on Aristotelian principles. This suggests a relatively optimistic view of the prospects for assessing virtue in efficient ways in schools, but the variety of virtues and contexts in which they are expressed cuts in the opposite direction.

If we could be sure that these component dispositions always develop and operate in unison as a tight bundle (i.e. co-vary), however people are raised and taught, then assessing virtue might be fairly easy. We could infer the whole of virtue from a part, if we had
a good measure of the quality of moral judgment (judgment of what it is best to do in particular situations), or moral perception (perception of the ethically significant aspects of specific situations), or emotional response to situations and actions. But we cannot be sure about this. Depending on our purposes, we might need several measures that get at different component dispositions. The goal in measuring virtue would be to gauge not just whether people have good ideas about what to do, but whether they are all together well equipped and disposed to act well. Some virtues pertain specifically to acting well in the face of such perturbing factors as fear, stress, and temptation, and in assessing virtue, we would need to know how well a person’s perceptions, judgments, and acts remain unperturbed when those factors are present.

Success in measuring virtue will depend, to a large extent, on being able to assess the different components or facets of virtue as they operate in concert. We have listed these components above, more or less, and would identify them more explicitly by saying that the attribute ‘virtue’ consists in the alignment of the following abilities and propensities:

- Acute moral perception (being able to see and distinguish morally important features in a situation);
- Appropriate moral emotion (having the right emotional response to the situation);
- Correct moral belief and reasoning (knowing or being able to work out what is appropriate and best to do in the situation);
- Active moral motivation (being motivated to do what one determines is the appropriate and best thing and to persist in seeing one’s action through).

According to Kristjánsson (2002: chapters 2 and 3) there is a further aspect of virtue, consisting of a propensity to comport or conduct oneself in an appropriate manner. The individual moral virtues (such as courage, honesty, and justice) would themselves be constellations of these various components, and their combined presence or absence would constitute a person’s state of character.

Much more could be said of the nature of virtue and of specific virtues, but with these basics in hand, we turn now to the theory of measurement.

**What is measurement?**

Addressing the question of whether virtue can be measured is no more likely to be fruitful without a clear conception of what measurement is than it would be without a clear conception of what virtue is. From a standpoint outside of psychometrics (the science of measuring mental attributes) and metrology (the science of measurement), it is natural to assume that there is a single unproblematic ‘scientific’ conception of measurement resting on uncontroversial assumptions about the nature of what is measured. This assumption is mistaken. The differences between understandings of measurement in psychometrics and in the physical sciences are significant for the question at issue, especially with regard to assumptions about the nature of what is measured.

Whether or not measurement of attributes requires that their structures be inherently additive or quantitative, or merely that their structures warrant ordinal ranking of ‘more’
or ‘less’ presence or strength of the attribute (making estimations of ratios based on cardinality and an absolute zero impossible), there is no doubting that the attributes must exist independently of attempts to measure them. The measurement of virtue presupposes metaphysical realism about states of character. Such realism has been contested by proponents of situationism (Doris, 2002; Harman, 1999; Vranus, 2009), but defended by others, including Snow (2010), Kristjánsson (2013), Fowers (this issue) and Jayawickreme et al. (this issue). We concur with these latter authors and regard both the situationist arguments and the social psychology on which they have relied as flawed, but we will set that debate aside in favor of matters less widely discussed. The first of these matters is that metaphysical realism about virtues and other such traits would preserve the common intuition that they are explanatory. It is fundamental to human affairs that we care not simply about outward conduct or behavior, but about its causes. Does an instance of harmful behavior arise from ill-will, a defect of character, non-culpable ignorance of the circumstances of action, impaired cognitive function, or something else? Treating virtues as explanatory precludes identifying them with the very patterns of behavior to be explained.

Yet, as Stephen Norris and his collaborators (2004) noted a decade ago,

The language of testing, especially of high stake testing, remains firmly in the realm of ‘behaviors’, ‘performance’, and ‘competency’ defined in terms of behaviors, test items, or observations. The validation models for high stakes tests frequently are founded on the concept of sampling from a population of behaviors or performances and, through statistical generalization, making inferences about what individuals can do. (p. 284)

When states of mind or intellectual virtues, such as understanding, are operationally reduced to patterns of behavior, such statistical inferences to what individuals ‘can do’ are, more precisely, inferences from behaviors elicited by test items to the wider patterns of behavior that should be observable in other contexts. Norris, Leighton, and Phillips argue that this is a ‘fatally flawed’ model for psychometric inference, and they defend an alternative model, based on the idea of an inference to the best explanation. In a psychometric context, the basic idea of this model is that a good test of a student’s understanding of photosynthesis, for instance, is one for which a good score could only (or best) be explained by the examinee’s possession of such understanding and a bad score could only (or best) be explained by the examinee lacking such understanding. If the understanding of a topic has distinct aspects that might or might not be present, it has a structure that is measurable, at least to the extent of permitting some ordinal quantification of difference between an examinee’s understanding and a model of full understanding (Elgin, 2004; Curren, 2006).

A recent, authoritative account of modern test theory (McDonald, 2013) suggests that not much has changed during the intervening years. McDonald recognizes that an ‘attribute is not “operationally” defined by just the set of items chosen to measure it’, given ‘our freedom to shorten or lengthen a test measuring a given attribute’ (p. 130). Yet, rather than abandon the misguided notion that the specification of a measure could fully define an attribute, McDonald offers an elegant but metaphysically bankrupt ‘save’:

The possibility of shortening or lengthening a test for an attribute rests on an idealization. In effect, we suppose that the items written come from a quasi-infinite set of items that would, if
written and administered, define and measure the attribute precisely. Such a quasi-infinite set has been called a behavior domain, or a universe of content. I prefer to call it an item domain.

(p. 130)

Explaining that current psychometric opinion favors the idea that there is just one form of validity, referred to as ‘construct validity’, predicated on all forms of available evidence that a ‘score is an acceptable measure of a specified attribute’ (p. 134), McDonald goes on to say that

If we regard the attribute as what is perfectly measured by a test of infinite length, then a measure of validity can be taken to be the correlation between the total test score and the true, domain score. The redundant qualifier ‘construct’ can be omitted. (p. 134)

Reliability, validity, and generalizability all come down to this correlation between test score and (quasi-infinite) domain score. Personal attributes are measurable by fiat, on this theory, since they are ‘precisely’ defined as nothing more than performances on (quasi-infinite) tests. Whether or not this escapes Norris, Leighton, and Phillips’ objections to the inferential logic of statistical generalization from a sampling of behaviors to a larger population of behaviors, it evades the fundamental questions at stake concerning the measurability of virtue. By contrast with classical and modern test theory, item response theory (IRT) seems to preserve a metaphysically or psychologically meaningful distinction between unobservable (latent) variables and their manifestation in observable behavior. IRT preserves the idea of construct validity, or measurement of an independently existing attribute, on the basis of ‘a theory of what the construct in question is and how it relates to other variables’ (De Ayala, 2013: 147).

The relationship between measurement and definition assumed by McDonald is a characteristic outgrowth of the way operationalism has been understood in psychology. Yet, the operational definition of psychological attributes does not require that they be equated with patterns of behavior, and reductionistic definitions of this kind were expressly opposed by Percy Bridgman, who first expounded operationalism in his 1927 book, The Logic of Modern Physics (see Chang, 2009). It was the embrace of Bridgman’s work on operations and measurement by logical positivists that engendered the long association between behaviorism and the operationalizing of psychological constructs as patterns of behavior. So to insist on a non-reductive realism about virtues and virtue measurement is not to argue that attempts to define psychological attributes operationally should be abandoned, merely that they be appropriate to the nature of the attributes and not assume that the specification of a measure could fully define an attribute.

There are good reasons to overcome the sharp opposition between two views of the empirical significance or content of theoretical terms: the positivist view that all meaningful terms must be fully operationally defined, and the Quinean view that all theoretical constructs have empirical content only through the empirical significance or testable implications of the larger theory of which they are parts. A more adequate view is that the empirical richness of theories and their component explanatory resources (constructs, attributes, laws) are enhanced by the number and variety of operations and measures by which they are anchored (Chang, 2004, 2009). To say this is not to rule out the usefulness of theoretical terms that are not operationally defined and play mediating roles between
other theoretical terms. Nor is it to insist that good operational definitions do or should fully define the terms or concepts, or that in psychology the definitions should be behavioral.

In sum, the nature of virtue as a multi-faceted attribute should be respected, but it is surely possible to do that and also develop measurement instruments through which the idea of virtue can be given well-defined empirical significance for the purposes of addressing it in a scientifically meaningful way.

The conventions of psychometrics allow that measurement may involve assignments of numbers on the strength of relationships of ‘more’ or ‘less’. By this standard, evaluation involving ordinal rankings of distance from an ideal will constitute measurement. Scoring essays using a holistic 5-point scale and model answers for each score would qualify. So too would the DIT, which is predicated on a developmental sequence of worse to better moral judgment schemas, orderable on an ordinal scale. These are both essentially forms of graded evaluation with respect to an ideal, and they qualify as measures by the standards of psychometrics, provided there are reliably discriminable degrees of difference or distance from the ideal. Accepting this understanding of measurement makes intuitive sense in the present context, inasmuch as some such estimations of comparative distance from an ideal of virtue seem feasible, while measures of virtue based on cardinal relationships might imply the counterintuitive possibility of meaningfully expressing comparative judgments of virtue as ratios or percentages (as noted in the ‘Introduction’ section).

One implication of this brief foray into the theory of measurement in psychology is that the construct validity of measures of virtue will not be well established until we have a satisfactory realist and non-reductive model of the structure and function of virtues – ‘a theory of what the construct in question is and how it relates to other variables’ (De Ayala, 2013: 147). Work on the development of measures can use as a point of departure a specification of the aspects of virtue, such as the one that concludes the previous section of this article, but the identification of these as aspects of a disposition or dispositional cluster is not psychologically well-defined (as noted above; see also Fowers, this issue).

Another important and useful lesson of this brief consideration of measurement theory is that psychometrically legitimate measures of virtue need not be machine-scoreable. Free-response test items, such as essays, are fair game, and might have some advantages, if cost and reliability of scoring can be managed – as they are in the Educational Testing Service’s AP essays and were in the California Golden State Examinations (Curren, 2004). The most obvious advantage of free-response items is that the one thing forced-choice items clearly cannot measure as well as free-response items is generative capacity. Forced-choice items supply at least some of what would be self-generated in the real-life moral responses of persons to situations, such as a limited menu of possibly relevant aspects of situations and possibly related principles or reasons for acting one way rather than another. Virtue requires a measure of moral imagination, or ability to bring to mind relevant possibilities and considerations. Even the perception of morally salient features of situations is generative, in the sense that one brings to mind for oneself the noticing of what matters. Free-response items might also be described on this basis as measures that are more authentic in the sense that they more closely resemble the tasks people face in being virtuous in life.
A final consequence of the foregoing is that even if being virtuous were a categorical attribute, and not even amenable to graded evaluation, the virtuousness of a population of students might nevertheless be measurable with enough accuracy for the purposes of program evaluation. An attribute whose possession by an individual is not additive in nature might nonetheless be present additively in a population comprising such individuals. Being amenable to quantification on the level of a population, virtue could be the object of useful measurement for the purposes of program evaluation, even if this were not the case for other educational purposes.

**The contexts and purposes of virtue assessment**

Having addressed the nature of moral virtues and measurement and drawn some general conclusions about the measurability of virtue, we turn now to matters of context and purpose. What are the presumptive contexts of, and purposes for, evaluating and measuring virtue, and how would these shape the legitimacy, methods, and likely success of evaluation and measurement? Judging the goodness of persons is so important to human social functioning that it is unavoidable, in fact, and to some extent automatic and unconscious. It may be as automatic as reading faces or as minutely studied as a Jane Austin novel, and how we go about it depends on our purposes and the constraints imposed by time, settings, resources, and the forms of contact, information, activity, and relationships involved. It will be useful to note some aspects of informal, out-of-school judgments of persons’ qualities, including moral virtues, before considering the prospects for conducting more formal assessments in schools.

**Everyday judgments of virtue**

We register the emotions that flash across others’ faces in our own emotions, without ever having been taught to do so and often without being able to pinpoint what has caused us to feel what we feel. Something of the character of a person is observable in this way in the relationships between the emotions discerned, desires inferred, and the features of the context that appear to be salient. Subtle signs of aggression, deception, or ill-will induce unease and mistrust, emotions we generally do well to heed, though our learned dispositions and judgments may get in the way. Expectations, implicit bias, and stress can also blind us to good qualities and signs of cooperation. We can read others’ minds in their faces, and it is in our nature to do so, but we do not always get it right and facial displays of emotion can flash across faces too quickly to be consciously parsed in real time. This illustrates two things: that estimations of character are commonplace, but also that the enterprise succeeds some of the time and fails at other times. Judging other people’s character from their faces, manner, words, and actions in everyday contexts is natural but often hard, and matters are much the same when it comes to judging others’ character in a more studied fashion – by getting to know them well over time. While one may have another person’s character pegged after long acquaintance, one may, equally, remain in the dark about the other’s character even after long acquaintance.

Recorded and slowed down, the patterns of facial movements that display different emotions can be identified by observers trained in Facial Action Coding (Ekman, 1995;
Ekman and Friesen, 1978). The Facial Action Coding System (FACS), developed by Paul Ekman and Wallace Friesen, is based on the classification of 43 distinct human facial muscular movements and identification of about 3000 combinations of those distinct movements as displays of emotions. Many of these combinations of movements are difficult or impossible to fully suppress or make voluntarily, so reliably tracking their occurrence might have evidential value in discerning gradations of virtue, since appropriate patterns of emotional response to situations are one aspect of virtue. If human emotional responses to actual situations and simulations are sufficiently similar, one could imagine simulation games enacted with facial monitoring and coding being one form of measure of a person’s goodness.

Readings of faces and body language are in any case often conjoined with inferences from more or less extensive histories of verbal and non-verbal conduct, in more or less varying circumstances, witnessed by one to many observers who are more or less virtuous themselves and more or less intimately acquainted with the person being judged. In all of these respects, more is better, and taking the measure of moral virtues is not unlike taking the measure of other personal qualities. There are qualities we value and are more or less able to discern in our neighbors, friends, partners, children, and the workers, professionals, and leaders on whom we rely. It would be surprising if there were not some reliability, but also limitations and variability, in our ability to judge the presence of those qualities. What is revealed in conduct and affect may not be equally evident to all observers, any more than gradations of particular talents and performances that display them would be equally evident to all observers. Experts in the domains of talent or excellence to be judged are presumably better judges or evaluators of the relevant forms of goodness, but judgments about who is good in general and who is good at one thing or another are often entangled in the evaluations made in employment and civic contexts. Knowing who will reliably perform well in expert tasks and who will reliably be honest may both be important, and the latter is not obviously harder than the former. For instance, in judging ability in a hiring decision, we would give some weight to test performance, if there is a relevant test, but in projecting likely performance we might rely more on the testimony of past supervisors and the degree of consistency evident in the academic record and other arenas of responsibility, much as we would in judging honesty.

The observation about better and worse judges has significant implications for character education and assessment. On the one hand, it shrinks the perceived distance between evaluating moral virtue and evaluating other forms of goodness. On the other hand, it suggests we must tread carefully in conceptualizing the scope and methods of character education and their relationship to the expertise required of character educators and evaluators. Expertise on substantive moral doctrine is not what is needed, and claims of such expertise should be met with skepticism, but we are inclined to think that qualities of ethical discernment and thoughtfulness are needed, along with skill in facilitating ethically reflective discussion and knowledge of moral development, learning, motivation, and the like.

**Evaluating character in schools**

If by ‘measuring’ virtue one has in mind the use of standardized instruments for quantifying the extent to which individual students possess moral virtues and make progress over
time in acquiring those virtues, then the idea of measuring virtues warrants careful scrutiny. One must ask what the purpose of such measurement would be and how the results would be used.

**High-stakes testing.** We would like to start by dismissing the notion of extending the recent enthusiasm for high-stakes testing and accountability schemes into the realm of virtues. The idea that it is productive to reward, penalize, and motivate teachers and school leaders on the basis of their students’ standardized test scores is ill-conceived, and it has proven to be counterproductive in practice. The notion seems to be that without such accountability schemes teachers are not sufficiently motivated in their work, and that being more highly motivated will improve their teaching. Motivational psychologists disagree, observing that attempts to ‘incentivize’ performance can yield heightened but also dysfunctional motivation (Deci and Ryan, 2012). Research on the effects of imposing such controlling structures on teachers indicates that it displaces the intrinsic motivation teachers bring to their work and induces anxiety that undermines their performance. They become more controlling and anxious in their interactions with students, and frame the value of schoolwork in more instrumental terms, with the result that students are less motivated and learn less (Pelletier and Sharp, 2009; Ryan and Weinstein, 2009; Vansteenkiste et al., 2009). This is only one of several detrimental aspects of testing and accountability schemes, but one that has special relevance for virtues. If one’s interest is in cultivating virtues in children, then testing their acquisition of virtues as a basis for judging their teachers is one of the worst things one could do.

Virtue involves an attachment to and pursuit of what is good because it is good. **Attachment or internalization** of values that yields healthy self-regulation, or fully integrated motivation, occurs when learners’ needs for mutually affirming relationships, competence, and autonomy are satisfied and they ‘understand and accept the real importance [of something] for themselves’ or have ‘identified with [its value] for themselves’ (Deci and Ryan, 2012: 89). In order for students to fully accept the value of something for themselves, they need an ‘autonomy supportive’ context that allows them to consider reasons without pressure to conform (Deci et al., 1994: 124). The evidence suggests that high-stakes testing of student virtues would directly undermine the social conditions in schools that are foundational to students’ virtue acquisition and embrace of the inherent value of the goods with which virtues are concerned.

What other purposes might there be for measuring virtues in schools? One likely answer is that measuring virtues is essential to evaluating programs of character education. Another answer might be that character education should be like any other form of education in having a student evaluation component, conceived as formative, summative, or both. A third answer is that measures of student goodness more systematic than those routinely used in schools could be usefully employed in decisions about school and classroom management: decisions about how to distribute students between different classes in the coming year, how to respond to disruptive behavior, and so on. We will speak to matters of program evaluation and routine student evaluation in the concluding sections that follow.

**Program evaluation.** No one should be field testing educational programs (or structural reforms), including programs in virtue education, without a basis in prior research and
tested theory. The body of theory and research on motivation and contextual factors favorable to fully integrated internalization of values is one such basis for program design, as are the program evaluation studies through which possible components of programs have been found to be efficacious. There are reasons why it is nevertheless preferable that new programs should be field tested. We will limit our remarks about this to observations suggested by the preceding sections.

The limitations of available methods for evaluating virtue commend a combination of pre- and post-intervention measures, and the measures chosen need not be ‘student-level significant’ measures designed to provide profiles of individual students’ degree of virtue acquisition. The goal in program evaluation is to establish the efficacy of a program, not to compare students with one another or with themselves over time, and the quality of information about programs may be enhanced through sampling strategies that do not attempt to learn the same things about every student. So, for instance, one might randomly select some students to discuss the ethical climate of their schools in focus groups, some before the intervention and others after, and code the discussion for frequency of salient normative terms.4 One might enlist selected classes on a similar pre- and post-intervention basis in writing essays on life plans (see Little, this issue) or the traits valued in friends. Other written measures might be random samplings of student work in any aspect of the curriculum in which virtue-related learning interventions are introduced – a matter addressed below. The progress of student ethical attunement and judgment might be tracked using a combination of such methods, and independent observations of student conduct and affect might be useful in establishing that the progress is not just cognitive. The forms of conduct that would have evidential value could be quite diverse, as Nicholas Emler (1996) noted some years ago:

To the extent that moral education enhances the skill and capacity of the individual to make moral judgments and produce moral argument its effects may be seen not so much in a change in the honesty, self-control, or altruism of that individual – the traditional behavioural criteria – as in the effectiveness of that individual as . . . a moral judge of others, as critic, persuader, advocate, provider of moral leadership. These effects should be most apparent . . . in terms of impact measured at the group or community level . . . [in] rates for truancy, exclusions from school, various kinds of victimization, damage to school property . . . participation in community service . . . (p. 123)

Evaluation of student ethical learning. There are ways in which we already evaluate students’ acquisition of virtues in schools, where the goals of educating them include not just skills and understanding, but commitment to certain goods – goods of inquiry, artistry, and the like – and consistency in pursuing and achieving those goods. ‘Effort’ figures significantly in grades, and ‘effort’ is a matter of striving toward the right things. This is not to say that the striving and commitment are moral striving and commitment to what is ethically valuable, but the task of evaluation in the two cases might be very similar. The question is where specifically moral virtues would lodge in the curriculum or extra-curriculum of schools in such a way as to provide a basis and vehicle for evaluating aspects of students’ acquisition of moral virtues. Apart from tests of what were once taken to be fixed, native abilities, what we test students on is what we have been teaching them.
So how would we teach moral virtue so that what is acquired could be put to the test? We offer first some cautionary observations about teaching and measuring virtues like courage and moderation in schools.

The extent to which virtues like courage and moderation can be acquired and measured in schools is surely limited by the limitations of context inherent to schools. How much supervised practice in moderation in the face of tempting pleasures, or courage in the face of fearsome threats, are we prepared to offer students in schools, and how many corresponding contexts for the expression of those virtues are we prepared to test them in? When it comes to those virtues, we might plausibly test understanding of them and opportunistically observe them in action — or not in action — when tempting and dangerous things that are not supposed to be present in schools are present in schools. We would not be able to test pivotal cognitive aspects of those virtues, such as the accuracy of perceptions of danger in the heat of the moment. We may be able to infer from students’ conduct in school and their academic success that they are able to defer gratification or work hard even in the face of many tempting things.

The idea of administering tests is that students will be prompted to respond to stimuli of the school’s choosing at predetermined times, and those stimuli and responses will usually be essentially verbal in nature and without real consequences. We can use practical tests in evaluating skill in laboratory procedures, musical and athletic performances, and other practical arts, but there are ethical and practical barriers to using such tests to assess virtues.

Notice that these virtues do not look much like the complex skills to which Julia Annas (2011) has helpfully compared virtues. What Annas does not quite say is that the focus of guided habituation in her account of virtue acquisition is on the learner deciding what to do for herself. Situations in life are complex in ways relevantly similar to the complexities faced by those with complex skills in the context of using those skills. The teacher must engage the student in recognizing and responding appropriately to the factors that are relevant. So the practice is substantially practice in recognizing and weighing the factors relevant to decisions and making those decisions, the rest being development of technique, as in violin performance or medical procedures. In order to test what students have practiced, then, we would — on an understanding of Aristotelian habituation that takes the role of good judgment seriously — test ethical discernment and judgment.

The analogy between general moral learning and learning the ethics of a sphere of professional practice is instructive. That is where we will begin, describing a form of Aristotelian habituation in ethical medical decision making. We will then briefly address three related forms of guided practice in ethical reflection, discernment, and choice and the forms of student evaluation appropriate to them. Our view is that any of these or similar forms of evaluation might play roles in evaluating programs in virtue-focused character education.

Example 1: Ethical coaching in a school of medicine. If one were designing an Aristotelian approach to teaching ethics in medical schools, its centerpiece would be supervised practice in ethical decision making (Shaw, 2011). Students would participate in the monitoring of cases, and when they meet as a class they would discuss what they and others take to be ethically salient in the cases they have observed. The instructor would facilitate

Example 1: Ethical coaching in a school of medicine. If one were designing an Aristotelian approach to teaching ethics in medical schools, its centerpiece would be supervised practice in ethical decision making (Shaw, 2011). Students would participate in the monitoring of cases, and when they meet as a class they would discuss what they and others take to be ethically salient in the cases they have observed. The instructor would facilitate
discussion, casting the students in the role of ethical consultants, and coach them as they think through the cases. The students would practice discernment, listening, and modes of consultation and ethical reflection conducive to making good decisions. This would occur against a background of prior instruction in a code of professional ethics and the basis for that code in the fundamental goods at stake in medical practice. The goal would be to cultivate professional integrity, and evaluation might best occur through longitudinal monitoring of the quality of patient care, patient satisfaction, and frequency of malpractice lawsuits. In the context of the class, the most authentic measure of learning might take the form of essays in response to case scenarios, scored for their efficiency in identifying and assessing the significance of ethically salient features and considerations. An alternative would be an oral ‘consultation’ constructed and scored on the same principles.

Example 2: The Promoting Alternative THinking Strategies curriculum. A starting point for character education is helping children become more attuned to the emotional dynamics of social interactions and more in the habit of thinking before they act – thinking specifically about the social and emotional dynamics of situations they may face and the likely consequences of different choices. The Promoting Alternative Thinking Strategies (PATHS) curriculum, a self-described program in social and emotional learning, was designed to do just this. It uses simple pictures of children in social transactions as a basis for teacher-facilitated discussions of what the children in the pictures are feeling, might do, and should expect to happen in response. It was validated through an observational counting method, but the learning stimulated by discussions of the pictures could presumably be evaluated by using novel pictures as prompts for free responses, scored for quality of noticing and understanding.

Example 3: Critical thinking projects. In 1996, Curren and a colleague launched an internship program that allowed college students who had taken classes in critical thinking and philosophy of education to spend a semester as teaching interns in urban elementary schools. They directed a variety of critical thinking projects that involved coaching 9- and 10-year-old students in producing reasoned essays and sometimes staging debates, often with a focus on decisions the children faced in their own lives. The preferred model was for children to identify for themselves the ethically significant questions they wished to address. One example of this was the question, ‘Should I join a gang?’ It is plausible that the children’s ability and inclination to think through personal decisions in light of relevant ethical considerations was strengthened by such learning. The ability to engage in such thinking is what was practiced and coached, and the essays and debate performances – sometimes witnessed by the entire school community – were the basis for evaluating their learning. Supposing those were scored for quality of ethical discernment, cogency, seriousness of purpose, and persistence, energy, and enthusiasm of engagement, the quality of performance in such essay writing and debates might be some of the best evidence of progress in acquiring good character one could hope for in the short term in a school setting.

Example 4: High School Ethics Bowl. A final example, more formalized in its prompts and scoring system is the High School Ethics Bowl, an off-shoot of the Collegiate Ethics Bowl. Both are forms of staged competitions in which student teams respond to
unexpected questions concerning cases scenarios they have previously had time to research and analyze. The questions that are most salient for our purposes are ones that pertain to the decisions of specific characters in the cases. The scoring is again focused on the quality of identification of ethically salient features of cases (completeness, emphasis, etc.) and cogency of the analysis and defense of an answer to the question. The Ethics Bowl is now the basis for ‘experiential learning’ on many college campuses in the United States and a growing number of high school campuses. It is arguably a model for cultivating and testing seriousness and thoughtfulness about ethical matters, though a measure of a person’s state of character as such it is not.

Conclusion

If evidence comes to light confirming the Aristotelian conception of habituation as producing a causally linked dispositional cluster of desire, affect, perception, reason responsiveness, and conduct, we may someday have a grounded theoretical basis for believing that measures of ethical perception and judgment of the kinds just described are more adequate as measures of virtue than we currently have reason to believe. Our understanding of virtue as a psychological construct might also advance in ways that provide a more illuminating basis for measuring states of character than we have at present. In the meantime, we should focus on evaluating forms of student learning as best we can, and evaluate character education programs through mixed methods and in light of our best theoretical understanding and knowledge of the relationships between meeting children’s needs and enabling them to be good and live well.

Acknowledgements

We are grateful to Kristján Kristjánsson for numerous formative conversations, and to James Arthur for his invitation to Curren to present the conference keynote address on which this article is based.

Funding

This work was made possible through the support of a grant from the John Templeton Foundation. The views expressed are the authors’ and do not necessarily reflect those of the Foundation.

Notes

1. In the physical sciences, ‘measurement is the attempt to estimate the ratio between two instances of a quantitative attribute’ (Michell, 2005: 287).
2. We use the term ‘aspect-score’ to refer to scores on measures of distinct aspects of virtue. We detail the nature of these aspects in the section that follows.
3. The term ‘disposition’ is a placeholder, in the absence of a psychological model of the motivational, affective, perceptual, and cognitive moving parts of states of character. Referring to a state of virtue as a disposition or cluster of dispositions is intended to indicate that a person’s moral qualities are not things she can turn on and off at will, or choose to make use of 1 day and not another. A state of character is understood to be intimately related to a person’s identity that persists through time.
4. On the methodology of scoring frequency of occurrence of moral terms in samples of writing or speech, see Frimer (2014).
6. See http://nhseb.unc.edu for cases, rules, scoring guides and other details of the National High School Ethics Bowl.

References


Jayawickreme E, Meindl P, Helzer EG, et al. (IN PRESS) Virtuous states and virtuous traits: How the empirical evidence of broad traits saves virtue ethics from the situationist critique. *Theory and Research in Education* 12(3).


Kristjánsson K (forthcoming) Some recent work on measuring virtue for character education: A critical overview.


Shaw MH (2011) *Coaching as a form of instruction and component of medical ethics education*. PhD Dissertation, Warner School, University of Rochester. Available at: https://urresearch.rochester.edu/viewContributorPage.action?personNameId=4321


**Author biographies**

Randall Curren is Professor and Chair of Philosophy and Professor of Education (secondary) at the University of Rochester (New York), Chair of Moral and Virtue Education in the Jubilee Centre for Character and Virtues at the University of Birmingham (United Kingdom), and Professor in the Royal Institute of Philosophy (London). His work in ethics, philosophy of education, and educational assessment has been supported by grants and fellowships from the National Endowment for
the Humanities, National Science Foundation, Spencer Foundation, Andrew Mellon Foundation, and Institute for Advanced Study in Princeton, New Jersey.

**Ben Kozee** is Lecturer in the Jubilee Centre for Character and Virtues at the University of Birmingham (United Kingdom). At Birmingham, he works on ethics and expertise in the professions (with a specific focus on the medical profession). He is the editor of *Education and the Growth of Knowledge: Perspectives from Social and Virtue Epistemology* (Wiley-Blackwell, 2014) and is the assistant editor of the *Journal of Philosophy of Education.*