Tipping the Scales: Social Justice and Educational Measurement

Tipping the scales: social justice and educational measurement

Zachary A. Stein

Dissertation Committee:
Dr. Catherine Elgin
Dr. Howard Gardner
Dr. Theo Dawson

A thesis presented to the faculty of the Harvard Graduate School of Education in partial fulfillment of the requirements for the degree of Doctor of Education.

2014

For Meghan, forever.

Table of contents:

Abstract: In this work I address foundational concerns at the interface of educational measurement and social justice. Following John Rawls's philosophical methods, I build and justify an ethical framework for guiding practices involving educational measurement. This framework demonstrates that educational measurement is critical to insuring, or inhibiting, just educational arrangements. It also clarifies a principled distinction between efficiency-oriented testing and justice-oriented testing. In order to explore the feasibility and utility of this proposed framework, I employ it to analyze several historical case studies that exemplify the ethical issues related to testing: (1) the widespread use of IQ-style testing in schools during the early decades of the 20th century; (2) the founding of the Educational Testing Service; and (3) the recent history of test-based accountability associated with No Child Left Behind. I conclude with a set of speculative design principles and arguments in favor of radically democratic school reforms, which address how the future of testing might be shaped to ensure justice for all.

Preface: the personal is political

> ...it is also the case that a sense of justice is continuous with the love of humanity.
>
> -John Rawls (1971 p. 417)

This is in many ways an intensely personal and emotional work, despite my attempts to maintain an impersonal and measured academic tone throughout. I am what is often referred to as a "high-achieving dyslexic," which means that I have been in a unique position with regards to standardized testing since I was about eight years old. In third grade I was given a series of standardized tests to determine the nature and extent of my disability and to justify my placement in special education programs. This was a profoundly important moment in my life. It is also an example of the social justice benefits that can come from testing. Through testing I was correctly identified as needing a different set of educational experiences, and I was fortunate enough to be in a school system that could provide them.

However, as a learning disabled student, I was also exposed to the injustices and biases of standardized testing. My performances on standardized tests were universally pitiful, and I felt very strongly that they did not reflect my capabilities at all. In my junior year of high school, when confronted with the high-stakes culture created by the test-based college admissions process, I wrote a letter of protest and circulated it, from the local superintendent of schools on down. This letter questioned both the idea of education as a competitive zero-sum game and the monopoly of the Educational Testing Service over the future of every student in my high school. Responses from administrators to this letter were varied, ranging from those who suggested

disciplinary action be taken against me to those who simply shrugged it off and told me that this was just the way it had to be—that I should simply work hard and deal with it. But a small number of teachers said I should try to do something about it. They said I should get my self into a position to make a difference. These teachers changed my life.

This leads to a further reason that what follows is personal: I love the American public educational system and have benefitted tremendously from it. I worry that recently the emphasis on high-stakes testing has been ruining a system that has done so much for so many. Public school teachers have meant the world to me: from the Dimmick sisters and Mr. Bourne at McGinn Elementary School, to Mr. Mealy and Mr. Good at Curtis Middle School, and Mr. Ray and the Plotts at Lincoln-Sudbury Regional High School. So while much of this work reads as a critique of US public school policy and practice, I am aiming to diagnose problems in order begin working toward designing a better system, one based on the ideals that have guided American schools since their inception: democracy, equality, and publicness.

This work is also personal because, over a decade ago, I co-founded an organization dedicated to the reform of standardized testing through the application of the learning sciences and philosophy. This non-profit, Lectica, Inc., is the brainchild of Theo Dawson, another of my truly amazing teachers, whose impact on my thinking has been so significant that she might reasonably be listed as a co-author of this work (indeed, there are sections in the Conclusion that stem from our joint publications). I put my best efforts toward the success of Lectica for over 10 years and continue to serve on the board of directors. This experience gave me an inside look at the testing industry and the role of testing in schools. Conversations with school leaders, teachers, and researchers exposed me to the assumptions that dominate the discourse about testing in schools as well as the intricacies and distortions of the testing industry. The arguments

presented here grew out of those years of struggling to change the system; they have especially grown out of my frequent frustrations with how testing is traditionally thought about and practiced.

And finally, this is a personal work because it was conceived, researched, and written during a time when I was serving as a caregiver, first for my mother and then for my wife. I learned to cultivate the alchemical process of converting pain and anger into reasonable and productive arguments. I channeled a great deal of emotion into this work, which has been peppered with tears and written in the shadow of profound illness and suffering. Along the way, I learned that love's labors, while unpaid, are infinitely rewarding. I am thrilled to be able to share this completed work with them both—they are well on their way to recovery.

I have incurred many intellectual debts (and some financial) while working on this project. So I must offer some thanks. At Hampshire College (where SAT scores are not factored into admission decisions) I began my work in philosophy under the guidance of two gifted teachers: Mario D'Amato and Robert Meagher. At the Harvard Graduate School of Education, Kurt Fischer and Catherine Elgin exceeded all expectations as teachers and academic advisors, and none of this would have happened without them. Quite literally, in fact: it was Kurt who convinced the Harvard admissions committee to ignore my GRE scores, which were so low my transcript was not even being considered (an irony that is not lost on me as I write this preface to my completed dissertation). Kate, on the other hand, provided support and shelter for my philosophical work in an academic climate where philosophy is little valued and less understood. Her unfailing insistence that my work was important and her demonstration and embodiment of the power of philosophy meant more to me than I can say.

My teachers, friends, and colleagues at Harvard and beyond have served as invaluable sounding boards as well as providing resources, advice, and relief form the solitary monotonies of writing. A short list includes: Clint Fuhs, Katie Heikkinen, Ken Wilber, Marc Gafni, Howard Gardner, David Rose, Veronica Boix-Mansilla, Mike Connell, Mary Helen Immordino-Yang, Denny Blodget, Marc Schwartz, Les Smith, Edward West, Tom Murray, Michael Herrick, Brian Hogan, Hans Despain, Sean Esbjörn-Hargens, Ben Williams, Dustin DiPerna, Jeff Carreira, Nick Hedlund-deWitt, Jennifer Worden, Amy Briggs, Topher Hunt, Zach VanRossum, Carol Bennett-Dessureau, Lauren Tenney, Rob Lindsley, Adrienne Tierney, Peter Blake, Johanna Christodoulou, Luke Shore, Tim O'Brian, Mike Hogan, John Reams, and Mark Fischler, and Metta McGarvy. I have also learned from all my students over the years, who probably taught me more than I taught them.

A few more specific thanks are needed: My wife, Meghan Byrnes copy edited and commented first drafts of the first two chapters. They are readable largely due to her efforts. My dear old friend from Hampshire, David Bowen, was distracted from writing science fiction by copyediting the entire first draft, which was absolutely invaluable. Sam Roberts took time away from revolutionizing the energy industry to read a later version and offered important feedback. Andie Poile provided helpful editorial insights on certain sections of later drafts while serving as a visiting scholar at the Mind and Life Institute. And, finally, there are numerous individuals with whom I've exchanged emails and ideas, evidence of which can be found in this draft: thanks to all my disparate and generous interlocutors! Of course, all weaknesses in the argument, errors, or inaccuracies are my own fault.

Introduction: social justice, the philosophy of education, and standardized testing

> The more any quantitative social indicator (or even some qualitative indicator) is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.... Achievement tests may well be valuable indicators of general school achievement under conditions of normal teaching aimed at general competence. But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways.
>
> -*Campbell's law* (Donald T. Campbell, 1979)

Many contemporary educational reformers argue that schools are the site of *injustices*. These reformers argue that by bringing in more money and better technology and improving school cultures, all students will get the education to which they are entitled. Moreover, these achievements will be reflected in better test scores, the proof that justice has been administered. Indeed, there have been recent cases where charter school start-ups or public schools under new leadership have yielded remarkable gains in student achievement, where test scores have been held high as proof of a school community's success in overcoming educational injustices. These trends reflect what has been one of the dominant narratives in educational testing for nearly a century: through the use of objective measurement we can better administer justice as well as make the educational system more scientific. The following chapters are, in part, an argument

that this idea is important and correct. However, what follows is also an argument that the same objective tests that are necessary for administering justice can become instruments of injustice when they are misused. This critical and cautionary argument is the main focus in what follows. It is important to understand the possible (and actual) injustices that can (and have) resulted from testing, because the tools of choice used by reformers today have the potential to become one of the main causes of injustice in the system, if they have not already assumed this unfortunate status.

Historically, some of the most profound educational injustices ever committed have involved the use of testing infrastructures in the wholesale reorganization of schools. Even a casual observer of the contemporary scene should be struck by the fact that testing is the only educational topic that grabs front-page panic-inducing headlines (aside from school shootings). The centrality of testing to the experience of American schooling is one of the keys to its importance as a social justice issue. Standardized testing has always been ideologically saddled with the task of administering justice in the schools. As explained below, this is why the 'craze' for testing continues and why testing continues to be touted by those who reform schools in the name of social justice.

Testing infrastructures continue to expand their reach and impact because they are built in light of a powerful half-truth. It is true that justice requires objectivity, efficiency, and accountability, even if only to assure that everyone is getting a fair amount of what goods there are to go around. All forms of objective measurement, including testing, can serve an important role in the administration of justice, potentially helping to assure the fair treatment of all who are impacted by their use. Below, these important values (objectivity, efficiency, accountability) are described as *true but partial* insights that emanate from standardized testing.

However, profound injustices can occur in any society if institutionalized rules and practices are used to systematically discriminate or misallocate what rightfully belongs to everyone, such as basic rights and freedoms. Social practices must be *designed* to treat everyone fairly; they do not naturally evolve toward justice. Tools like the scales in marketplaces and the ballot boxes used during elections are needed to help administer justice. But as dishonest vendors and election fraud show, the same tools that are intended to promote fairness can also be used to deceive, manipulate, and subjugate. Measurement tools and practices, including tests, can easily come to serve as instruments of injustice, perpetuating distortions of social practice so that benefits accrue to those with power or are otherwise unfairly distributed. It is this seeming paradox—that testing can be both just and unjust—which serves as the starting place for the philosophical work undertaken here.

This introduction provides an overview of the key philosophical themes treated in this work, serving as a model or miniature of the whole, with the additional task of justifying the project and its methods. The introduction begins with a discussion of justice and the philosophical methods used to think through what justice is and to adjudicate between what is just and unjust. While many philosophical and ethical ideas are discussed, focus throughout this work is placed primarily on the methods and models of the great moral philosopher, John Rawls. Narrowing the focus to a single thinker is intended to simplify the discussion of social justice—drawing on the value of an exemplary theory instead of trying to draw on the value of a set of theories or constructing a whole new theory of justice.

There are, in fact, many theories of justice that could be used to clarify the relationships between standardized testing and social justice along the lines undertaken here. Habermas, Nussbaum, Dewey, and Freire all come to mind. No doubt, taking these different theories as a

starting point would result in an argument different from the one developed here. Yet it is hard to imagine (given the continuities of these 'competing' theories of justice) that there would not be an "overlapping consensus" on many of the central insights. In any case, the goal here is to demonstrate as clearly as possible the social justice issues implicated in standardized testing, not to engage in philosophical debates about competing theories of justice. Rawls is taken as a primary guide not because his is the only or best theory, but because, for the purposes of this study, his views provide the minimal complexity necessary to deal with the issues that are of primary concern in addressing the relationships between testing and social justice. As a result of this near singular focus, this work serves as an interesting, if unconventional, introduction for educators to the central ideas of one of the twentieth century's greatest moral philosophers. But it is not merely an exegesis of Rawls. This work is an original application of his ideas in the domain of educational theory and the philosophy of education.

It is also worth noting that many of the issues that occupy attention in contemporary academic discourses about social justice and education do not receive explicit attention here. Race, class, gender, ableism, and a host of other issues are truly pressing social justice concerns in contemporary schools. However, analyzing the relationships between testing and these issues must wait until after the groundwork laid here has been set. Without a general framework and an understanding of the history of testing, it is not possible to adequately address these more complex topics. Indeed, one of the most pressing projects needed to advance future work on justice and testing is a careful and critical analysis of the differential impact of testing infrastructures on various racial and socio-economic groups.

Likewise, international perspectives on social justice and standardized testing in countries other than the US are not discussed here. The perspectives and history offered here are distinctly

American (the term American and US are use interchangeably throughout, as per typical conventions; this is not intended to slight the "other" geographical areas rightly called American). Standardized testing infrastructures have been designed and implemented differently in different countries. But they have had (and are now increasingly having) profound impacts on social justice and education throughout the world. Applying the framework built here in the analysis of international testing practices is an important next step.

Finally, and again because of the need to start with the basics, there is no explicit discussion of *the political economy of the testing industry*. Since very early in its history, testing has been a large and profitable industry in the US. The testing industry represents the first real successful commercialization of psychology (not counting phrenology as psychology), and it continues to be one of the most lucrative and impactful fields in applied psychology (Brown, 1992). In fact, today the testing industry (which includes not only test makers, but also test-prep companies, computer vendors, and a variety of data-driven educational consultancies) is larger and more profitable than ever. This is thanks to federal legislation that has ensured that nearly every child in the US is tested multiple times during their years in school. Moreover, given the largely for-profit nature of the testing industry, the political economy of testing today involves public money going into private hands on a massive scale. These companies face the traditional pressure to maximize returns to shareholders and thus shape research and development efforts accordingly, increasing the intensity of trends toward automation and achieving economies of scale. Again, these important issues require separate treatment, building on the initial framework offered here.

Before beginning an overview of the work, a few words are needed about Rawls's philosophical methods, which are emulated throughout. This brief overview of Rawlsian methods is followed by a thumbnail sketch—a miniature—of the work as a whole.

Methods: reflective equilibrium, provisional justifiability, and the need to make sense of history

Ethical frameworks, like the theory of just educational measurement being built here, cannot be directly confirmed or disconfirmed in the way scientific hypotheses can. But there are systematic methods for building and justifying them. Rawls and others (Daniels, 1996) argue for a methodological approach to building and justifying ethical frameworks, which aims to make them suitable for guiding reform and policy. The components of ethical frameworks—principles, judgments, and empirical generalizations—must be explicated and then "tested" against the varied experiences and competing accounts already available on the topic. The ethical framework thus undergoes a process of iterative revision until it is brought into a state of broad reflective equilibrium (Rawls, 2001). This is a state of "provisional justifiability," occurring when an ethical framework is shown upon reflection to be internally coherent, empirically tenable, and consistent with considered moral experience. Justifying an ethical framework thus requires demonstrating its ability to maintain a broad reflective equilibrium, its ability to handle a wide variety of particular cases while still maintaining its logical consistency, its ability to account for accepted facts, and its capacity to make sense of our most assured moral judgments.

To clarify, reasonable individuals typically work to achieve a narrow reflective equilibrium whenever a novel occurrence forces them to reconsider their beliefs, a process by which long held beliefs are opened to revision in light of new experience. For example, someone believing that standardized testing practices promote discipline, higher standards, and

accountability is likely to revise this belief, or at least qualify it, when confronted with the details of large-scale faculty-organized cheating in urban school districts (discussed in Chapter 5). Revising this belief requires revising other beliefs that are related to it, but not necessarily letting go of commitments to discipline, standards, and accountability. On an individual level, reflective equilibrium is the process through which ethical reasoning leads to learning and conceptual change, as the integration of new experiences reshapes existing beliefs (Habermas, 1990; Kohlberg, 1984).

Philosophers work to achieve a broad reflective equilibrium. This is a process during which philosophical principles and judgments are systematically "tested" against the best of our knowledge and the various convictions and realities of our lived experience. For example, a philosophical principle that would exclude all standardized testing from educational practices (e.g., "categorizing students is unethical") must be revised, or at least qualified, if it is to reflectively accommodate arguments and data concerning the use of diagnostics with special populations, the fair organization of large-scale social benefit programs, or the importance of advancing the learning sciences. Revising this principle would have ramifications throughout the problematized ethical framework, as demands for internal coherence and consistency set off a cascade of conceptual revisions. For an ethical framework to maintain a broad reflective equilibrium it must be in dynamic contact with, and learn from, a variety of potentially dis-equilibrating realities, difficult case studies, and provocative thought experiments. Philosophers have used this method to build and justify a variety of ethical frameworks, most recently for bioethics (Buchanan, Brock, Daniels, & Wikler, 2000), disability advocacy (Nussbaum, 2006), and international law (Hayden, 2002).

An ethical framework concerning educational assessment must be able to account for our considered judgments about a wide variety of testing practices. Therefore in Chapters 4 and 5 the ethical framework developed in Chapters 1 through 3 is used to analyze a set of exemplary case studies that are rich with ethical complexities: the IQ testing practices in the 1920s, the national testing infrastructure built by the Educational Testing Service (ETS) in the 1950's, and the ascendency of test-based accountability in the twenty-first century. Organizing the available historical materials in a consistent way, specifically around testing practices and their related justifications, scaffolds the systematic application (and evaluation) of the proposed ethical framework, showing it to be reflectively equilibrated, and thus demonstrating its provisional justifiability. These historical sections also provide an overview of some of the most important episodes in the history of testing in the US.

Social justice and educational measurement: a thumbnail sketch

John Rawls's (1971; 1996) philosophical methods also involve the use of complex *representational devices*, which can be thought of as structured models or thought experiments. More broadly, model-based reasoning in the sciences involves the deployment of a variety of "useful fictions" that simplify phenomena and exemplify the properties or processes of interest (Elgin, 1996). One of the most common kinds of models is the *miniature*, such as the scale-models used in engineering and systems biology that represent large structures or long timelines in ways that "shrink" them down so they can be seen at a glance. The rest of this introduction is just such a miniature; it aims to bring a large and complex work into view by shrinking it down and distilling its most important properties. This means stepping back from the particular arguments to gain a view of the whole.

The simplest way to miniaturize my overall argument is to consider the design of standardized testing infrastructures as if from behind a Rawlsian "veil of ignorance." The "original position" is the central representational device deployed by Rawls. It is intended to clarify the objectivity and universality of the "moral point of view." The original position is basically a set of decision-making constraints that support reasoning about the nature of justice; it simply asks us to consider the basic institutions of a society as if ignorant of our eventual place in them. The archetypal case is drafting a constitution without knowledge of who or where you will be in the society it creates. From the perspective the original position it would be irrational to draft a constitution supporting slavery or limiting voting rights to landowning males because there is no guarantee you would not end up enslaved or disenfranchised. Engaging Rawls's thought experiment means that instead of viewing social structures from my perspective (i.e., that of a well-educated white male), I am forced to consider society from everyone else's perspective as well (e.g., that of a woman, of a minority, etc.). A social system that can be viewed as reasonable from this meta-perspective is one that provides justice for all.

It is worth quoting Rawls (2001, pp. 14-17) at length summarizing the motives and design of his famous thought experiment:

> We start with the organizing idea of [a just] society as a fair system of cooperation between free and equal persons. Immediately the question arises as to how the fair terms of cooperation are specified.... They [are to be] settled by an agreement reached by free and equal citizens engaged in cooperation, and made in view of what they regard as their reciprocal advantage, or good.... The difficulty then is this: we must specify a point of view from which a fair agreement between

> free and equal persons can be reached. This point of view must be removed from and not distorted by the particular features and circumstances of the existing basic structures [of society]. The "original position," with the feature I have called the "veil of ignorance," specifies this point of view. In the original position, the parties are not allowed to know the social positions or the particular comprehensive doctrines [worldviews] of the persons they represent. They also do not know persons' race and ethic group, sex, or various native endowments such as strength and intelligence.... We express these limits on information figuratively by saying the parties are behind a veil of ignorance.... The significance of the original position lies in the fact that it is a device of representation or, alternatively, a thought-experiment for the purpose of public- and self-clarification....

This is not the place to get into the complexities surrounding the original position and its various formulations (see: Freedman, 2007). Rawls intended this thought experiment for use only in adjudicating between different philosophical principles of justice, and thus not for thinking about more specific social structures and institutions. However, for the purpose of miniaturizing the overall argument, the thought experiment serves as a valuable heuristic and allows us to cut directly to the chase.

The overarching themes of this work can be distilled into a single question: what kind of standardized testing infrastructure could be agreed to in the original position? This question is at the center of the *theory of just educational measurement* that is the overall focus of this work. Of course, there is much more to it than that. As will be explained shortly, a theory of just

educational measurement requires related theories about the nature of institutionalized measurement in general, as well as a supplemental philosophy of education, and a system of distinctions concerning the dynamics of testing-intensive educational reforms, specifically a system centering on the difference between justice-oriented testing and efficiency-oriented testing. But before adding these details to the miniature being built here, it is worth exploring the simplest way of thinking about testing infrastructures and social justice.

Several issues are clarified immediately through the "ideal role-taking" exercise of imagining that one could end up anywhere in the systems affected by testing infrastructures. Key stakeholder groups emerge, each with their own institutionalized relationship to testing and each containing a range of individuals (from least well off to most well off). Students and their parents are one group, and their range can be viewed both in terms of socio-economic factors and in terms of learning abilities. Teachers are another group, again including a range of individuals who vary according to their skills and socio-economic positions. Then there are the administrators at various levels within the school system (from principal to superintendent) who are also differentially positioned. Policy makers and politicians constitute another group, as do psychometricians and others representing the interests of testing companies. There are of course other stakeholders (e.g., educational researchers, college admissions officers, etc.), but the heart of the argument resides with these main groups, including students and their parents, teachers, administrators, policy-makers, and test providers.

The most vulnerable individuals in the social structures created by testing are the least-well-off students and their teachers (e.g., learning-disabled students in an inner city school and their special education teachers); the least vulnerable are the politicians and those representing the interests of testing companies (e.g., Arne Duncan and Educational Testing Service

executives). As in many cases where injustices occur, the most vulnerable—those who are potentially most seriously affected—are also the least empowered and the furthest away from influence over the systems that profoundly shape their lives. The task of building a just standardized testing infrastructure (as with any basic social structure viewed from the original position) requires taking as primary the perspectives of the least well off because this is the social position that is of greatest concern from behind the veil of ignorance (e.g., it is the place you would least like to end up in the system). Justice requires assuring benefits to the least well off while attempting to maximize overall fairness within the system.

Broadly speaking this means a testing infrastructure that rewards those who are already advantaged while punishing those who are already disadvantaged is unjust. The distribution of benefits resulting from such a testing infrastructure simply further disadvantages the least well off. As discussed below, the history of testing from the early IQ testing movement to recent policies supporting test-based accountability has mostly fit this unjust pattern of differential reward and punishment.

Social justice is also implicated at the level of test design and administration. For example, objectivity and standardization are required by justice—this is the moment of truth expressed by those who tout the social justice benefits of testing. Indeed, many of the most egregious cases of injustice due to testing have involved a *pretense* of objectivity that disguises the existence and impact of overt biases and errors in test design and scoring.

As discussed below, all individuals have a right to objective measurement. On the other hand, even truly objective tests can create injustices, especially when they are used in certain kinds of high-stakes contexts or focus only on a narrow range of constructs and item formats. Beyond objectivity, individuals have a right to measurement practices that are relevant and

beneficial. Arguments from the original position suggest the plausibility of these metrological rights, but a full understanding requires bringing in the rest of the theory of just educational measurement. The remainder of this introduction is a summary of the overall argument and introduces the key conceptual distinctions in broad brushstrokes—key terms and ideas are introduced here that will be elaborated in the body of the work. Rawls provides the basic conceptual elements throughout, beginning with an account of the relations between justice and institutionalized measurement in general.

Social justice and institutionalized measurement

Chapter 1 deals with the ancient relationship between measurement infrastructures and social justice. This discussion is necessary because standardized testing is best understood in light of the long history of institutionalized measurement. Standardized testing constitutes a recent and ongoing chapter in the socio-political history of quantitative objectivity. Chapter 1 revolves around Rawls's (1971) three principles of justice, which can be thought of as ethical design principles for the basic structures of a society. A society's basic structures are those institutions and legal codes, such as its taxation mechanisms, judiciary processes, and educational system, which constitute the society as a shared social world. Individuals participate in these basic structures as a result of participating in society at all; and because these basic structures shape the lives of every person they are the subject of theories of justice.

Measurement infrastructures are basic structures. In fact, measurement practices were some of the first basic structures to be perceived in terms of their impact on social justice. From the systems of scales and bushels used in ancient marketplaces to the precisely calibrated instruments used in modern science, measurement has always been a crucial component of

shared social life, literally structuring and facilitating mutual understanding and consensus. Because measurement infrastructures are basic structures Rawls has lessons to teach about the way they ought to be designed.

Rawls's principles of justice state that a society's basic structures should be designed so that all individuals are granted the same rights and freedoms, all inequalities result from conditions of fair opportunity, and the distribution of unequal benefits always advances the position of the least well off. The lesson of the unjust miller (an historical character introduced in Chapter 1)—who changes the size of the village bushel by *fiat* to serve his own needs—teaches that objective measurement is a prerequisite for social justice. The creation of the metric system in Revolutionary France, which corrected this kind of injustice, was as much an ethical undertaking as a scientific one.

But objectivity is not the only key to a just measurement infrastructure. Tracing the spread of the metric system and the advance of large-scale measurement-intensive organizational structures reveals that objectivity and its accouterments (e.g., scientific technologies and expertise) can create injustices. The lesson of the unjust bureaucracy—which unilaterally imposes objective measures that reshape the practices of individuals and groups—teaches that measurement infrastructures must also be relevant and beneficial to those who are most affected by them. The arguments in Chapter 1 clarify three principles of just institutionalized measurement, which follow from Rawls's principles of justice: all individuals have a right to (1) objective measurement whenever possible and preferable, (2) measurement practices that are relevant to their needs, and (3) measurement practices that are beneficial. But these are general principles that apply to all measurement infrastructures. For use here in addressing issues

involved with standardized testing these general principles must be supplemented by a theory about what makes for a just educational system, for which Rawls again can provide inspiration.

S*ocial justice and education*

Chapter 2 distills a minimal philosophy of education from the wide-ranging but sparse reflections on education that Rawls offers in his major works. The first insight Rawls offers concerning the philosophy of education is about the nature and allocation of *educational primary goods*. These are the educational experiences that all individuals are entitled to—the "amount" of education owed to everyone. More specifically, educational primary goods are defined as those educational experiences that, under normal circumstances, can be reasonably believed to reliably provide for the skills and dispositions that enable full participation in a society's civic culture and public sphere, as well as those that enable individuals to pursue a self-chosen conception of the good life.

Related to the idea that there are certain educational goods owed to all is the idea that schools ought to function as part of a system of institutions that secure fair equality of opportunity. Importantly, Rawls does not shoulder the educational system alone with the task of securing equality of opportunity, but instead positions schools in relation to both economic and political institutions. This complex system of institutions, in which schools play a critical role, is intended to provide for a kind of 'pure procedural justice' in the allocation of opportunities.

Finally, school systems in a just society must provide for the possibility of self-actualization. This idea follows from a key concept in Rawls's moral psychology, known as the *Aristotelian Principle*. It posits that all individuals have an inborn preference for learning and

exercising increasingly complex skills in contexts over which they have control (e.g., a preference for non-alienated learning and labor).

These are the three lessons Rawls teaches about the nature of just educational institutions: (1) they provide for educational primary goods, (2) contribute to a system of institutions that secures equality of opportunity for all, and (3) make possible individual self-actualization. When this minimal philosophy of education is combined with the aforementioned principles of just institutionalized measurement the basic outlines of a theory of just educational measurement come into view.

Social justice and educational measurement

Chapter 3 lays out the central components of a theory of just educational measurement, which begins by clarifying the differences between physical and psychological measurement. Tests are value-laden in ways that physical measures are not. Tests necessarily create interpersonal relationships in which there are both epistemic and social power differentials—the one who gives the test is, by definition, both more knowledgeable and more powerful than the one taking it. Power differentials of this kind are not necessarily a part of physical measurement practices, which involve people but are not primarily or inherently *about* people. In other words, treating testing practices as if they were physical measurement practices ignores the power relationships inherent in testing. Nonetheless, analogies to medical diagnostics and engineering (both of which involve the use of physical measures) have been present throughout the history of testing (Brown, 1992). Today the most common form of this conflation is referred to as *the education commodity proposition*.

The education commodity proposition is the simple yet powerful idea that education can be treated like any other commodity. This turns educational measurement into a means for putting a number on the value of educational processes, which can then be converted into monetary terms, usually in the form of cost-benefit analysis. The classic statement of this ubiquitous idea is: *how much education are we getting for our tax dollars?* Testing is seen as a necessary part of answering this kind of question and thus for monitoring changes in the financial value of educational processes. The reasoning is that if you cannot measure it then you cannot monetize it, and if you cannot monetize it then you have no way of knowing if your investments have paid off. This is the basic dilemma facing those who invest in educational institutions and then must prove return on investment (e.g., governments, philanthropies, venture capitalists). This same way of thinking impacts administrators, teachers, students, and parents, all of whom at different times and for different purposes deal with the financial meaning of test results—and in doing so, run the risk of reducing the value of education to the terms of the education commodity proposition.

Importantly, this testing-enabled representation of educational value is not wrong; it is *true but partial*. Economic efficiency is necessary for any viable enterprise. Indeed, efficiency-oriented testing in general is necessary for the maintenance of most large-scale educational institutions. Efficiency-oriented testing transcends but includes the education commodity proposition, contextualizing it in terms of broader social goals, usually codified in terms of educational standards. Efficiency is always determined relative to some goal, and efficiency-oriented testing is always carried out relative to some definition of what education ought to be. Social justice concerns arise when the quest for objectivity and efficiency, in themselves necessary and good, results in testing infrastructures that distort social relationships within the

schools by limiting what counts as good education to what can be measured by tests. Again this reasoning is simple yet powerful: efficiency is a non-controversial good for all schools to pursue, objective measurement is required for determining and improving school efficiency, so therefore only that which can be objectively measured can be included in considerations about what makes a school good.

Because they are designed for system-level surveillance, efficiency-oriented testing infrastructures are often irrelevant and harmful to those most affected by them (e.g., students and teachers)—thus violating the second and third principles of just institutionalized measurement. Moreover, efficiency-oriented testing tends to undermine the possibility that schools can provide the kinds of educational primary goods justice requires, let alone foster fair equality of opportunity and individual self-actualization. So while much of the discourse about testing and justice has historically centered on bias, cheating, and the damage done when tests lack objectivity, most contemporary testing-induced injustices are, in fact, the result of an excess of objectivity, which is narrowly defined and mechanically implemented. In a sense efficiency-oriented testing provides too much of a good thing. The necessary and reasonable pursuit of objectivity and efficiency are taken too far, overriding the metrological rights of students and teachers while drastically truncating the scope of what is perceived as educationally valuable.

Justice-oriented testing infrastructures, on the other hand, are built to assure that objectivity and efficiency are achieved, but not at the expense of being relevant and beneficial to those most affected by them. A testing infrastructure that honors the metrological rights of students and teachers is more likely to actualize the commitment of an educational system to provide for the full range of educational primary goods. It contributes to a system of background institutions that promote equality of opportunity and facilitate the self-actualization of all

students. Needless to say, there has never been a pure instance of justice-oriented testing (just as there has never been a pure instance of efficiency-oriented testing). But attempts at justice-oriented testing have been in evidence since the first testing infrastructures were built in the early decades of the 20th century. Testing infrastructures have served a variety of functions that are a necessary part of any educational system committed to social justice, including administering a kind of pure procedural justice in the allocation of opportunity, assuring the equitable distribution of educational primary goods, and identifying the unique learning needs of each student.

But the differences between efficiency-oriented testing and justice-oriented testing cannot be grasped in the abstract. Examples from the history of testing must be examined to determine the worth of any theory of just educational measurement. This work thus deals with three historical case studies: (1) the origins of educational measurement [Chapter 4], (2) the founding of the Educational Testing Service [Chapter 5], and (3) the recent history of test-based accountability following in the wake of No Child Left Behind [Chapter 5]. The concluding chapter takes these lessons from history and ventures a guess at what preferable futures for testing might look like, arguing that justice-oriented testing remains a profound and important possibility, now more than ever.

1: Social justice and institutionalized measurement

> An unjust measure is an abomination to the Lord.
>
> Proverbs 11:1

> Inequality before law implies unequal laws or rights in relation to measures: some people decree them, others have to put up with them; everyone has a measure of his own, the strong imposing theirs on the weak. The measure is not impersonal but rather human; it belongs to some, it does not belong to others, and it is dependent upon the will of whoever has the power to enforce it.
>
> -Witold Kula (1986, p. 122)

Imagine a farmer who is taxed a certain number of bushels of grain depending on the size of his land, yet each year both the size of the bushel and the measures of land size change to meet the needs of the local magistrate. Or imagine a marketplace where different vendors use different scales for weighing the same materials and change scales depending on who is buying. Consider being in a local community that has used traditional measurement practices for centuries in the distribution of land, only to have a centralized government body enforce the use of newer scientific measures. In such cases, where land was once distributed according to a measure that combined both size *and* probable yield, it is now distributed according to a universal standard for determining area, which is indifferent to the local variations in soil quality that accounted for the stability of the traditional practices. Finally, think about having properties

of your own body and mind measured by officials who use the results to determine your eligibility for certain social benefits, yet these systems of measurement are demonstrably biased, scientifically specious, and susceptible to corruption.

These examples from history demonstrate the relationship between systems of institutionalized measurement and social justice. In post-industrial societies, universal standards for physical weights and measures are taken for granted—a meter or a liter is the same for everyone everywhere. But this is a relatively recent state of affairs, and it took centuries, thousands of scientists, and several political revolutions to bring this about (Kula, 1989; Tavernor, 2007). In fact, for the majority of humanity's civilized existence a debate has raged about the relationship between justice and measurement, creating an archetypal relation best exemplified by Lady Justice, who wears a blindfold and holds a scale, one of the most ancient instruments of objective measurement. Today, newly created measurement systems are rekindling deep concerns about the relationship between justice and measurement, especially in areas like biometrics, econometrics, and psychometrics—the focus of this work. The social justice issues raised by these new measurement systems have strong analogies to those raised in the past concerning our most basic systems for physical measurement. For this reason I begin building an ethical framework for educational measurement by considering ethical issues in *historical metrology*, which is the historical study of physical measures and measurement practices.

This chapter is an introduction to John Rawls's theory of justice by way of certain themes in historical metrology. Rawls offers a set of key concepts as a part of his ethical framework, specifically the idea of society's basic structures and the principles of justice that ought to guide their design. I will explain this ethical framework with reference to the history of measurement

practices in order to clarify the structure of problems at the interface of social justice and institutionalized measurement systems. This is preparation for the central argument that unfolds over the rest of the work, where I address the relationships between social justice and educational measurement in post-modern societies.

First I compile a set of observations and generalizations about the basic functions of measurement systems in society. I argue that they form an important part of any society's *basic structure*, playing a key role in the system of institutions that fundamentally shape social life. A society's basic structure determines the scope of its members' liberties and rights, as well as the distribution of the basic goods that result from social cooperation. According to Rawls, these basic structures are what theories of justice should be about. A theory of justice is a framework for determining how to build basic structures that are *fair*—structures that will create the kind of "background justice" that enables trust, equality, and autonomy. It is easy to see that measurement infrastructures are undoubtedly fit to be the focus of theories of justice. They form a part of basic structures that affect a wide variety of basic goods, including everything from food and money, to job opportunities, health care, and self-esteem. As historical metrology shows, justice and measurement have been mutually defining terms since the dawn of civilization.

Having clarified the role of measurement infrastructures as basic structures, I then explore what it means to consider measurement infrastructures in light of a broader theory of justice. Rawls offers a set of philosophical principles that clarify the nature of justice, defining it as fairness in the arrangement of basic structures. He proposes a framework for making decisions about the fairness of basic structures, providing a method for adjudicating between just and unjust arrangements. I demonstrate how this framework works, specifically demonstrating how it

can be used to survey the moral complexities of institutionalized measurement systems. Using examples from historical metrology, I show how Rawls's theory clarifies what constitutes *a just use of institutionalized measurement*. I bring these insights into Chapter 2, where I turn to focus on educational systems as basic structures, and begin building a Rawls-inspired framework addressing educational measurement.

Measurement infrastructures as basic structures of society

> Our topic...is that of social justice. For us the primary subject of justice is the basic structure of society, or more exactly, the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation.
>
> -John Rawls (1971, pp. 6)

> We enter into them by birth and exit only by death.... the institutions of the basic structure have deep and long term social effects and in fundamental ways shape citizens' character and aims, the kinds of persons they are and aspire to be.
>
> -John Rawls (1996, p. 68)

In complex societies social action is a highly coordinated affair, involving economic, legal, and political systems—a network of institutions that congeal into a *basic structure,* which sets the terms of social cooperation and distributes the advantages that result from it. According to Rawls, the basic structure of a society establishes the "background justice" that conditions and

shapes the lives of each member. Not every social structure is a part of the basic structure because not every institution has deep and pervasive effects on the shape of society. Basic structures are those that touch all members in some way, especially those that determine their access to basic rights and goods. Rawls argues that these structures should be the primary focus of ethical frameworks concerning social justice because they set the conditions in terms of which the actions of individuals, groups, and associations take place. No matter how free and fair a specific transaction appears to be, we cannot say it is just without understanding the broader social institutions in which it occurs. This was a lesson learned clearly in the segregated American south, where ostensibly un-coerced, fair transactions at "separate but equal" businesses and schools were, in fact, reinforcing unjust legal structures that grossly distorted all human relationships. "Thus we seem forced to start with an account of a just basic structure. It is as if the most important agreement is that which establishes the principles which govern this structure" (Rawls 1999, p. 257).

Rawls argues that *fairness* ought to govern the design of all basic structures. This is because basic structures are a non-negotiable precondition of life in a complex society. We do not opt in to or join up with society; we are always already members of it, and will always live in terms of at least some of its basic structures. Of course, individuals can emigrate between societies, but they cannot return to a state of nature; we participate in basic structures of one type or another from the day we are born until the day we die. We participate in them and conduct our lives according to them, yet we did not choose them—we just happened to be born in a certain time and place. They shape our fate as if they are a part of nature, yet they are social constructions (Searle, 1995). Therefore, the arrangement of these basic structures is of special ethical concern, especially their bearing on the life prospects of individuals belonging to different

groups, be those economic, generational, religious, or what have you. So while there are many acceptable ways to design institutions that one can freely choose to be a part of, such as a club (which charges for membership) or a scientific association (which excludes non-experts), when an institution is a part of the basic structure that *everyone* must participate in, different organizational principles must be applied. According to Rawls, justice must be the dominant design principle for institutions, such as legal systems, tax codes, and educational systems, that shape the very fabric of social life by structuring the terms of collaboration and the distribution of benefits.

Measurement practices were some of the first social institutions to function as basic structures and to become the subject of theories about social justice. Measurement practices exemplify what basic structures are and how they function to create the background justice of a society. Lessons from historical metrology will clarify the relation between measurement practices (as basic structures) and social justice. This sets the stage for an exploration of the ways in which educational measurement practices have many of the same social justice implications as physical ones, serving as basic structures, as a part of legal codes, and being implicated in the distribution of basic social goods.

It has been argued that the ancient Egyptians first invented standardized measures in tandem with geometry in order to redistribute land and levy taxes each year after the Nile's annual flood destroyed the previous year's plots. While the true origins of geometry may lie elsewhere, it is undoubtedly true that the first standardized measurement practices were invented in response to pressing social needs (Duncan, 1984). Measurement practices—involving measurement instruments and norms for their use—are as old as civilization itself and were some of the first social institutions ever established. Scales and measuring rods have been found

amongst pre-civilized humans, and all ancient civilizations had complex systems of weights and measures. Historical accounts tell of an astonishing diversity of pre-modern measurement practices, all of them built out of necessity and evolving in response to the needs of their creators. Peoples in regions where land was scarce had precise measures of area, whereas peoples in regions with abundant land had more approximate measures. Those who dealt in gold had complex systems of precise scales for trading, while those dealing in oats or hay had systems of bushels and baskets, no less complex, but certainly less precise. Most early measures were anthropocentric, often literally involving a part of the human body (such as a thumb or foot), and they were all created to serve cognitive, practical, and communicative needs (Kula, 1986).

Measurement systems emerged and have been institutionalized to address recurring social situations in which coordinated action depends upon the creation of a shared understanding of specific qualities and quantities in the objective world. Innovations in measurement stem from situations in our social life where it is necessary to achieve mutual understanding about a state of affairs that is repeatedly problematic, yet also *consistently objectively determinable*. These types of situations have a similar epistemological structure. They require multiple parties to be able to verify the amount or quality of some thing or things that concern them. Measurement systems are part of those social practices that require the reliable differential determination of objective traits in objects of concern. Considering even the most basic instance of measurement bears out this account.

If I measure a board by myself as I build a table, say by laying down the length of my arm from elbow to finger tips and marking it as one *ell*, through the act of measurement I have positioned the board in a space of meaning to which I ascribe universal intersubjective validity. Which is to say, in the practice of measuring the board I am, in effect, saying that *anybody and*

everybody who measured this board this way would find the same thing. Of course, my forearm is longer or shorter than yours. So as soon as one friend shows up to help me build, we are thrown into a negotiation about how precise an *ell* we need and whose arm it will be if a judgment must be made. This begins a process of refinement that, over the long run, results in a tape measure, marked in both metric and United States customary units, which any modern table builder would use off the shelf without a second thought. Usually the measures we agree to slip into the background and become part of our taken-for-granted measurement infrastructures, which are the condition for the possibility of a vast amount of highly coordinated social actions.

Measurement infrastructures form a part of society's basic structure because they shape social life in fundamental ways, specifically by providing a means for coordinating social action in relation to objective realities. Consider the measurement practices involved in scientific research, engineering, and economic exchange, or in the administration of basic governmental tasks, such as taxation. Measurement practices can facilitate these complex social activities because they provide a reliable index of reality that has been codified to consistently generate a broad consensus, ideally universal. Thus measurement infrastructures, like legal infrastructures, are both systems of knowledge and systems for guiding action and administering conduct (Habermas, 1996). They require knowledge about the invariant properties of objects and occurrences, which in turn entail the creation of instruments and practices that reliably differentially respond to those properties. Reflectively (sometimes scientifically) codified measurement practices then come to structure broad swaths of social life, often to the point of being woven into systems of law. Collections of measurement practices often congeal into an infrastructure and come to function as a taken-for-granted aspect of social life, so much so that unjust systems of measurement have been perpetuated for centuries due to sheer force of habit.

This was the case with some systems used to administer taxation under feudalism in medieval Europe, a process of social inertia often aided by the rule of law (Kula, 1986).

History shows that measurement infrastructures have been, and continue to be, invented and institutionalized for social uses and wedded to systems of law. Aristotle's research for his *Politics* included a comprehensive survey of the existing city-states' constitutions, and while the full fruits of his research were lost, what remains suggests that measurement practices, especially in the market and the field, were a major concern for those out to administer justice in the ancient world. It is no coincidence that the first ancient systems of measurement were accompanied by the first legal codes, or that legal systems and measurement systems have co-evolved since the dawn of civilization (Duncan, 1984). In fact, measurement practices were some of the first institutions to qualify as what Rawls would call basic structures of society.

In Aristotle's political anthropology, he recounts the societal need for *Metronomi* (commissioners of weights and measures), to be appointed by lot, and tasked with "seeing that sellers [in the market] use fair weights and measures" (*Constitution of Athens*, Ch. 51: see Ross, 1921). "There were also the Sitophylaces (corn commissioners)... who watched over prices and the weights of loaves [of bread sold at market], which they had the power to standardize" (Duncan, 1984 p. 13). The earliest legal systems functioned to establish, among other things, measurement infrastructures as basic structures by using the force of law to assure that specific measurement practices would be reliably regarded as the "true measure" (e.g., sometimes literally the king's *foot*). Like other basic structures, such as those for voting and jurisprudence (which are also the subject of ancient constitutions), the institutionalization of measurement practices serves to facilitate trust between strangers, mutual understanding at a distance, and fairness through the standardized treatment of cases. But this same constellation of law and

measurement can be, and has been, used to exploit, systematically discriminate, and fallaciously justify inequity.

Consider the result of different laws governing the jurisdiction and powers of ancient and medieval commissioners of weights and measures, who roamed the market places attempting to regulate the use of metrics, e.g., scales, rods, bushels, and loaves. Kula (1986) reports of wide variations between markets that were only miles apart, with ethnic and religious differences often resulting in different rights with respect to measurement. Wealthy merchant guilds found ways to change legal codes in order to allow them the power to regulate their own measures, which then inevitably changed in response to the needs of the guild. When yields are good, the bushel (or what ever is the standard unit of exchange) is big; when yields are poor then the bushel is small, yet the price of the bushel, its exchange value, remains unchanged (Kula, 1986 pp. 43-71). This kind of metric manipulation was common practice and remained a ubiquitous part of economic life for centuries. It is important to understand how *unit setting* differs from *price setting* as a means for merchants to offset unexpected problems with supply or production. This is an issue we will return to when we discuss the manipulation of educational measurement systems, where units (e.g., cut-off scores) are set differently in different places, or change in the same place from year to year, often in order to offset problems with "production" (e.g., lowering the cut-off score next year assures the appearance that our school is producing students as good as or better than last year's).

In modern economies, when yields are poor it is the *price* that goes up, but the unit (or measure) remains unchanged—a trend starkly exemplified by the worth of a gallon of gasoline, the size of which remains constant despite the highly variable and politicized nature of its cost. In ancient and medieval markets, it was more often the unit or measure that fluctuated in

response to problems with supply or production. The loaf of bread, a staple of ancient and medieval life in almost all of Europe, was a basic unit of exchange, structuring the access most city dwellers had to one of their most basic sources of nourishment. But the loaf was a notoriously unstandardized unit, fluctuating largely in response to the supply of grains, sometimes imperceptibly, sometimes enough to incite a riot. For this reason many local political authorities regulated the size of the loaf, to assure fairness, but also for other governmental reasons, such as to mitigate the risk of famine, build up supplies for war, or head off political upheavals (Kula, 1986 pp. 71-80). As discussed below, it was not until the metric system was spread via political revolution, colonialism, and scientism that many of our basic units of measure and exchange became "impersonal," ostensibly scientific, and generally perceived as fair—these measures and units thus became objective social facts that are now a part of the taken-for-granted infrastructures that facilitate coordinated social action.

Measurement was a political preoccupation in the ancient world and continued as such for centuries, with measurement infrastructures remaining a perennially contested subject in the expanding discourse about social justice leading up to the Enlightenment. While the role of measurement infrastructures in contributing to the background justice of a society was understood, and the ideal of *just measurement* had been codified in a variety of ancient and medieval texts (from constitutions to religious scripture), it would be centuries before many basic measurement infrastructures attained the universal and objective status they had always been counterfactually ascribed. It is no surprise then that many of the largest social and scientific undertakings in early modern history were focused on creating measurement infrastructures for science, industry, and government. The standardization of measures for temperature, distance, and weight dominated scientific discourse from the seventeenth through the nineteenth centuries,

resulting eventually in the establishment of international standards for measurement, which were understood as a precondition for scientific collaboration (Travernor, 2007). The rapid invention of agricultural technologies, mills, and eventually the steam engine necessitated the large-scale institutionalization of the standardized measures being created by scientists, many of whom were on the payroll of industrial benefactors (Porter, 1995). In the socio-political arena, the building of the nation-state entailed monumental undertakings in the design and construction of measurement infrastructures, for taxation, military technology and inscription, economic exchange, and many other measurement-intensive activities necessitated by modern governments (Scott, 1998).

Because of their unique and irreplaceable social functions—e.g., their role as basic structures—measurement infrastructures have been the focus of some of the most intense and prolonged political and scientific efforts in history (Alder, 2002). This has resulted in a world-historical process during which the multitudinous local measurement practices that had evolved naturally in all societies were systematically replaced by measures that were centralized, scientific, government-sanctioned, and internationally calibrated. The most famous instance of this was, of course, the institutionalization of the metric system in the wake of the French Revolution. In this case Enlightenment ideology posited that objective, standardized measurement was a primary means for facilitating both social justice and scientific progress. As one rallying call during the revolution put it, "...one law, one weight, one measure.... For all people, for all time" (Kula, 1986 pp. 267). These ideals led to the international proliferation of the metric system, and its eventual near universal adoption by all the nations on Earth. Of course, the complex history of this process is nothing like a linear story of progress in which the new "true" measures were welcomed by the people.

I have argued here that the earliest measurement practices emerged to meet a specific type of social need for objectively determined coordinated action. Because of the ubiquity of this need and its occurrence in situations that involve the allocation of basic goods such as food, land, and money, measurement practices quickly became understood as social institutions in need of regulation in order to assure social justice. That is, they became understood as basic structures of society. This new understanding of the relationship between measurement and justice sparked reform efforts, eventually aligning with emerging scientific practices, which resulted in a worldwide movement toward the standardization of measures and the creation of international measurement infrastructures.

This cursory and simplified history of physical measurement is intended only to demonstrate their role as basic structures. With this established, we can now turn to a discussion of Rawls's principles of justice, which are intended to regulate the design of basic structures. Historical metrology will again help illustrate philosophical ideas, leading ultimately to an understanding of what constitutes *the just use of institutionalized measurement*.

The just use of institutionalized measurement

> We believe that as a matter of principle each member of society has an inviolability founded on justice which even the welfare of everyone else cannot override, and that a loss of freedom for some is not made right by a greater sum of satisfactions enjoyed by many...
>
> -John Rawls, (1999, p. 131)

> Measures succeed to the degree they become "technologies of the soul." They provide legitimacy for administrative actions, in large part because they provide standards against which people judge themselves.... Measures succeed by giving direction to the very activities that are being measured. In this way individuals are made governable.... [Measures] create and can be compared with norms, which are the gentlest and yet most pervasive forms of power in modern democracies.
>
> -Theodore Porter (1995, p. 45)

As explained in the Introduction, Rawls employs a variety of philosophical methods to justify two principles of justice. He argues that these principles best characterize the design requirements for the basic structure of a democratic society conceived as *a fair system of cooperation between free and equal citizens*.[1] That is, these principles are intended to clarify and operationalize an ethical ideal (or moral point of view) according to which all members of a society are entitled to the same basic rights and liberties. The other great "fruit of the revolution" (i.e., aside from the metric system) is this modern ideal of social justice, as a society in which there are not first-class and second-class citizens, and in which political power is conceived as a function of popular sovereignty, not divine right.

Other contemporary philosophers have followed Rawls in attempting to unpack the ethical insights at the heart of modernity. Habermas (1990), in particular, has argued in favor of

[1] The terms 'citizen,' 'person,' and 'member of society' are used more or less interchangeably throughout this work. All the social justice issues discussed here concern *everyone*, not merely those who have a particular legal status in a certain country. In this way I take Rawls's principles of justice according to their most *universalistic* reading.

similar principles that specify (even more fundamentally) the ideals of interpersonal reciprocity and mutual respect that form the normative core of modern moral consciousness. What these approaches have in common is the goal of carefully constructing a set of statements—a conceptual or principled framework—that explicates the intuitive commitments that implicitly guide moral judgment. Ethics is thus conceived as a process of self-clarification, which attempts to articulate, justify, and bring coherence to our varied judgments and results in increased clarity and understanding (Freeman, 2007). Here we find the methodological ideal of an ethical framework in broad reflective equilibrium, which was discussed at length in the Introduction. Furthermore, once such a framework is built it can then be used to address difficult ethical problems, such as those encountered in specific political, economic, or social reform efforts.

Rawls articulates his two principles of justice as follows (2001, p. 42):

1. Each person has the same indefeasible claim to a fully adequate scheme of equal liberties, which scheme is compatible with the same scheme of liberties for all.
2. Justifiable social and economic inequalities must by definition satisfy two conditions: first, they must be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they must be of the greatest benefit to the least advantaged in society.

The first principle addresses the basic liberties and rights of members in a society, stating that all are entitled to the most adequate system of liberties possible, constrained only by the need to assure an equally adequate system for all. This is a fundamental idea in modern moral

and political philosophy. It can be found in Kant's notion of a "kingdom of ends," and more recently, in Habermas's notion of a "system of rights," both of which articulate the insight that, in a just society, "the liberty of *each* is supposed to be compatible with equal liberty for *all* in accordance with a universal law.... [All people] have a *right to the greatest possible measure of equal individual liberties*" (Habermas, 1996 pp.120-121, emphasis in the original). This means, among other things, that a society should not design its basic structures so as to circumscribe the liberty of some as a way to increase the liberty of others, as is the case with slavery or exploitative labor. As discussed below, measurement infrastructures, because of their role as basic structures, play a critical role in establishing (or undermining) conditions that enable this kind of reciprocal system of equal liberties. Moreover, in later chapters, it will be shown how educational measurement infrastructures—standardized tests—are likewise implicated in establishing (or undermining) this most fundamental form of background justice, specifically by structuring access to the basic goods needed for the exercise of liberties, such as self-respect and the skills and knowledge necessary to be an active participant in civic life.

The second principle is two fold, with the first part addressing how the roles and offices in society are assigned, and the second part addressing the distribution of the goods that result from social cooperation. Beyond the basis of equality and liberty established by the first principle, the first part of the second principle states that what social and economic inequalities there are must result from offices and positions that are open to all under conditions of fair and equal opportunity. This means that while inequalities between individuals in society do and often should exist (as is the case with talents and motivations), these inequalities should not be the result of an inequitable distribution of opportunities provided by a society's basic structures. For example, the complex divisions of labor that have characterized industrial and post-industrial

society create inequalities of wealth and power between individuals. These differences are unjust if aspects of a society's basic structures (e.g., its educational institutions) systematically give certain members more ready access to privileged positions in society than others. The second principle continues, stating that those inequalities that do result even from fair conditions of opportunity must always in some way contribute to the benefit of the least advantaged members of society. This idea—known as the "difference principle"—argues that a just society does not allow those with a greater share of goods to use their advantage solely to further their own interests. A just society is redistributionist to some extent and builds its institutions to assure that no benefits accrue to the "top" that do not also somehow improve conditions at the "bottom." As discussed below, this has major implications for the design of measurement infrastructures, from those that structure tax codes to those that structure educational systems.

These principles are listed in order of priority, with the first principle taking precedence over the second, and the first part of the second taking priority over the last. This means that considerations about the structure of social and economic inequalities, which are the focus of the second principle, are only to be addressed in contexts where the basic structures have already established a foundation of equal liberties for all. Simply put, it is a moot point, in terms of social justice, whether the allocation of jobs to slaves is done according to fair equality of opportunity. Likewise, considerations about the fair distribution of social goods, which is the focus of the second part of the second principle, are of little concern if the offices and positions in society are awarded unfairly. Again, simply put, it is a moot point whether or not a society's resources are fairly distributed if roles and careers open to some are not open to others, due, for instance, to institutionalized racism or sexism.

In the following sections, these principles of justice are explained in detail and their implications are elaborated. Lessons from historical metrology help illustrate the meaning and importance of these principles, while applying them to this historical material provides insights into what constitutes *the just use of institutionalized measurement.* These insights are brought forward into the following chapters, where educational measurement is discussed as an instance of the fundamental relationship between institutionalized measurement and social justice.

Measurement and the liberties of equal citizenship

> By the priority of liberty I mean the precedence of the principle of equal liberty over the second principle of justice.... The precedence of liberty means that liberty can be restricted only for the sake of liberty itself.
>
> -John Rawls (1971 p. 214)

> Every act of measurement is an act marked by the play of power relations.... Because local standards of measurement were tied to practical needs [and] were "an attribute of power and an instrument of asserting class privilege," and because they were "at the center of bitter class struggle," they represented a mind-boggling problem for statecraft.
>
> -James Scott (1998, pp. 28-29)

Rawls's theory of justice is a *liberal* conception, placing primary emphasis on the indefeasible right of each person to a fully adequate scheme of basic liberties. Liberalism is often seen as originating with John Locke's arguments about the origins of private property in self-ownership in a state of nature. Here all "men" (sic) are born free and equal with certain inalienable rights, which it is the duty of the state to protect. Rousseau likewise theorized about the basic right of all members of society to equal citizenship, offering a vision of democracy as deliberation among equal citizens concerning the common good. Beginning with the Reformation, and implicated in the great Wars of Religion as well as in the French and American Revolutions, liberalism is a complex tradition in political philosophy.

These original ideas lead to what Rawls (2000, pp. 366) has called the "liberalisms of happiness," such as those of David Hume and Adam Smith, which stress primarily economic rights of contract, trade, property, and consumption. Thus liberalism is often misunderstood as predominantly an economic doctrine, with an overreaching ideal of maximizing aggregate happiness. Rawls argues strongly against this tradition, "since the basic ideal [of the "liberalisms of happiness"] is that of maximizing happiness, it is a contingent matter whether doing this will secure the basic freedoms" (*ibid*.). For example, consider a socio-political arrangement that raised the standard of living for everyone (say, through government subsidies) by removing the right to an education for some subset of the least well-off, taking funds from their education and directing them towards creating menial jobs and consumer abundance. Rawls suggests that a utilitarianistic liberalism of happiness (and some forms of libertarianism) can lead to this kind of arrangement, where the promotion of aggregate happiness limits freedoms that promote equity and human dignity.

Rawls (*ibid*.) contrasts these "liberalisms of happiness" with the "liberalisms of freedom," stemming from the tradition that runs from Kant through Humboldt and the German Idealists. This tradition, in which Rawls places himself, stresses the ideal of free self-governing persons who aim to develop their humanity, pursue freely chosen conceptions of the good, and actively participate in shaping society. The main concern in this tradition is not with protecting the rights of individuals with regard to economic exchange and consumption (although this is seen as important), but with protecting the rights of individuals with regard to their participation in society as free and equal citizens. As with other "liberalisms of freedom," like Kant's and Habermas's, Rawls's theory begins with the structure of the moral personality, building out from a specific conception of moral agency toward the social structures that allow for it. According to these conceptions, the most basic liberties—those that must be protected above all others—are to be selected in light of an *ideal conception of the person*.

This "political conception of the person" is not a psychological description or prescription, but a philosophical construct. It is intended as an abstract and schematic representation of the person, one that embodies a philosophical idea, specifically the minimal person needed to sustain society as a fair system of cooperation between free and equal citizens. It requires only that individuals use their liberty in ways that do not impinge upon the rights of others to an equally adequate basic scheme of liberties. Likewise, it privileges no specific conception of the goods individuals pursue within the system of liberties they create. It is thus indifferent to "all reasonable comprehensive doctrines"—such as religions, philosophies of life, or overarching political or scientific worldviews—which all persons live by in some shape or form and that give meaning to cultural groups, historical epochs, and individuals' lives (Rawls,

1996).[2] This idealized citizen is similar to the idealized decision-makers in the original position discussed in the Introduction; they are free from content and thus serve as structures and scaffolds for our thinking about basic rights and liberties. Here the goal is to figure out what the basic structure needs to secure for all members—the basic "all-purpose" means necessary to secure their existence as free and equal citizens in perpetuity.

Rawls is specifically interested in securing the "liberties of equal citizenship," which are those freedoms that are essential social conditions for the full development and exercise of the "moral powers" individuals need to participate in society as free and equal citizens (1971, p. 197; 1996, p. 293). These include an individual's sense of justice and ability to pursue a self-chosen conception of the good. Thus liberties concerning private property, for example, are to be considered in terms of their impact on the development of a certain kind of person, specifically a person who understands what it means to be fair, who has the capacity to decide for themselves what is of most value in life, and who has the right to build the skills needed to pursue that life. A legal system that does not protect private property, argues Rawls, does not provide a basic structure conducive to the pursuit of self-chosen ends, nor does it allow for individuals to develop a sense of what it means to cooperate under a fair system of law. On the other hand, a system that allows for the indiscriminate and unchecked accumulation of private property (in wealth and other means of production) is unlikely to sustain a fair system of liberties between citizens, because of the disproportionate influence on public life that comes from radical inequalities of wealth. The "fully adequate scheme of equal liberties" described in the first

[2] To be clear: this ideal conception of the person is not indifferent to *all* comprehensive doctrines, only to all *reasonable* comprehensive doctrines. For example, a comprehensive doctrine that promotes terrorist acts against innocent people is clearly not reasonable. The same notion applies in discussions below where individuals are said to have a right to pursue any *reasonable* self-chosen conception of the good life. Rawls's specific definition of *reasonableness* is discussed in detail near the end of Chapter 3.

principle is a scheme that adequately maintains the conditions that foster an ideal of moral agency. This is the criterion according to which Rawls argues that some liberties should be seen as basic (such as freedom of thought), while other liberties should be seen as less essential, and in some cases, even as undermining justice (such as the freedom to accumulate large personal fortunes).

Consider, for example, cases from historical metrology in which basic structures were maintained that benefited a small but powerful elite while undermining the ability of the less fortunate to change their political and economic positions and self-understandings—a case where a system of liberties was collectively maintained that worked against the development of people's abilities to think and act as free and equal citizens. Based on court records and other archival evidence, Kula (1986) recounts the politics of measurement involved with the collection of feudal dues and taxes during the seventeenth and eighteenth centuries in Eastern Europe (also discussed in Scott 1998, p. 28). The measurement practices involved in administering these transactions were conducted by noble and clerical claimants. They involved traditional customs that needed to be maintained, such as the use of a certain number of bushels for collecting feudal rents paid in oats. So much was attached to these measurement customs—religious and superstitious meanings, in particular—that they were rarely changed. However, the inherited power of the nobles and clergy entitled them to adjust certain aspects, such as the size of the measure. They had the power, literally, to decide on the size of the bushel, while leaving the overall transaction (the *number* of bushels) ostensibly unchanged in the eyes of age-old tradition.

Rulers of different regions and of the same region at different times would become known for their measures, some being considered as having a just, fair, or true measure, while others were seen as manipulative and unjust, with the worst cases going to court, or causing

uprisings. “The local lord might surreptitiously or even boldly enlarge the size of the grain sacks accepted for milling (a monopoly of the domain lord) and reduce the size of sacks used for measuring out flour.... while formal custom governing feudal dues and wages would thus remain intact (requiring, for example, the same number of sacks of wheat from the harvest of a given holding), the actual transaction might increasingly favor the lord. Kula estimates that the size of the bushel used to collect the main feudal rent increased by one-third between 1674 and 1716...” (Scott, 1998 p. 28). These practices were widespread, significant, and a likely cause for the general popular support of the metric system as it spread throughout Europe.

Simple cases of *metrological injustice*—such as the unchecked power of a local lord to alter measurement practices—reveal the structure of injustices more generally. Importantly, deception and inaccuracy are not at issue in the case of the unjust miller, although they often do play a role in cases of metrological injustice. In the cases discussed by Kula, everyone was generally aware that the bushels and sacks changed in size. Construction techniques for bushels and sacks were a topic of much debate, as different materials and building practices lead to warping, shrinking, or expanding. The fact that there was some imprecision in measurement practices was expected, and in some cases embraced. These were, after all, farmers who worked fields measured by *stone throws* (e.g., by custom they determined the boundaries of their fields in informal and relatively inaccurate terms, such as making the distance a stone can be thrown or an arrow shot). The social justice issue here is not about whose measure is most accurate, it is about who has the *right*, the liberty, to impose *their* measurement practices on others, and why. Consider again the quote that began this chapter, “Inequality before law implies unequal laws or rights in relation to measures: some people decree them, others have to put up with them; everyone has a measure of his own, the strong imposing theirs on the weak” (Kula 1986, p.122).

Metrological injustices often operationalize an asymmetry or inequity in the design of a society's basic structures, especially its legal systems.

The farmers who brought their oats to the unjust miller were capable of objectively determining many things about their own crop, including relative volumes, weights, and aspects of its quality, always with a certain consistent degree of accuracy. They used their own measures as a major part of everyday life, typically using ancient customary measures that allowed for collaboration and trade in the local region. Bickering about bushels being heaped (rising above the rim) or shaken (to pack more in) was common, but the idea of a shared objective referent remained. Or at least the option remained to call off a negotiation about measures when someone was not listening to reason and looking at the scales. But when dealing with the feudal lord, who inherited control of both military power and the means of production (e.g., the mill), competent people with knowledge of measurement were stripped of their rights and liberties with regard to adjudicating and administering measurement practices. The metrological rights of the lords were backed by the traditional religious laws of the land, which instantiated the feudal hierarchy system.

Unequal status with regard to measures creates a rift in the social fabric, disallowing the participation of all as free and equal citizens. Measurement practices are basic structures affecting everyone, so Rawls would argue they must be designed in light of their impact on individuals' abilities to develop into free and equal citizens. A legal system that allows for key measures to be set and changed by *fiat* (where the power to define the measure assures its use) will likely violate Rawls's first principle of justice. Of course, at some level all measures are set by fiat in so far as a choice must be made between numerous valid ways of systematizing the measurement of a property such as, for example, length (as debates about the metric system

make abundantly clear). However, the issue here concerns unequal social status and power with regards to measurement: when measures are *manipulatable* by fiat, when they can be both set and changed according to the dictates of a (small) group of individuals who have been uniquely empowered, either legally or by tradition. When unaccountable powers control crucial measures we are likely to find basic structures that give more liberties to some at the expense of limiting the liberties of others.

Measurement practices facilitate social cooperation by attuning social practices to realities in the objective world. When used in a just manner, they allow groups to see the same thing in the same way, assuring that like cases are treated alike, while facilitating trust at a distance and freedom from bias. However, when one group possesses disproportionate power to manipulate measures and wields this power to the benefit of some and not others, it undercuts the conditions that enable measurement to function as a just social practice.

Individuals subjected to a "reign of false measures" are stripped of both their social power and their epistemic autonomy, that is, their power to make truth claims about the world. Consider the feudal-era farmers who, when given a new measuring rod by an unjust tax collector, would publicly mark it with the length of their traditional measure. This was done both as a sign of protest at the injustice of the dictate and as a way of keeping much needed epistemic continuity with existing buildings and fields, known for years in terms of a different rod. Unjust manipulations of measures turns an instrument for coordinating social activity with reality into an instrument for manipulating the perception of reality in order to control social activity. It alienates individuals from their power as both social actors and cognitive knowers, clearly undercutting the development of the broader powers of autonomous and collaborative free and

equal citizens. Moreover, because so many measures are so basic in the ways they structure social life, metrological injustices can be pervasive and systemic.

This reveals a first lesson about the just use of institutionalized measurement: *the right to objective measurement is a basic liberty of equal citizenship.* It has been shown that measurement practices function as basic structures in society, as instruments and norms for coordinating social action with the realities of the objective world. They are implicated in transactions of innumerable types and of wide-ranging significance, including the most basic aspects of economic exchange. Objective measurement (where possible and preferable to existing practices) is a basic right—a condition of equal citizenship and something that it is the legitimate function of the state to maintain. This is because free and equal citizens, to be such, cannot have their interactions systematically distorted, nor can they be disempowered in their understanding of reality, due to their inequality with respect to measures. Conversely, justice demands measurement practices that are (among other things) based on objective standards that are set publicly and in the context of reflective practice.

The right to objective measurement often entails a right to scientifically refined measurement instruments, although this is not always the case. For many purposes where measurement practices structure social life, scientific standardization is not necessary or useful. For example, in a situation where a fair distribution is sought for many individuals eating from one store of rice, it does not matter that the basket in which rice is distributed is exactly one metric quart, but it does matter that everyone use the same basket in the same way when they receive their share. In general, from a social justice perspective, it is more important that a measure be objective (demonstrably changeless regardless of context, content, and user bias) than that it be scientific (iteratively refined to be as accurate and objective as possible).

Consensus takes precedence over scientific accuracy in the adoption of measures, and it is only in some cases where scientific levels of accuracy are a condition of consensus. Objectivity is a complex philosophical concept, but only a minimal definition is needed to sustain the arguments offered here. Throughout this work 'objectivity' and 'objective measurement' are defined in terms of the qualities listed above: a practice is objective when it is demonstrably changeless regardless of context, content, and user bias.[3] Scientifically refined measures are often more objective, but it is possible to achieve objectivity using measures that are not rigorously calibrated. Historically, the science of measurement was never an important factor in the adoption and proliferation of measures, except in cases where it was scientists or engineers doing the adopting. Nevertheless, the science of measurement advanced in large part due to an emergent ethical and political ideal that each citizen has a right to objective measurement.

In the following chapters, it will be shown that educational measurement involves *the right to be objectively measured*. The lessons of the unjust miller also apply to schooling and the use of tests in sorting students for efficiency management, especially in the early days of IQ testing, where the lack of objectivity in testing created demonstrably unjust educational systems. But before discussing educational measurement, there are two more essential lessons to learn about the just use of institutionalized measurement more broadly. It is necessary to continue exploring the second principle of justice, because even a measurement infrastructure set up to provide for the basic right of objectivity can be a part of a basic structure administering injustice.

[3] For a detail discussion see the "Excursus on objectivity and the three principles of just institutionalized measurement" that concludes this chapter.

Measurement and distributive justice

> The rules of background institutions required by the two principles of justice are designed to achieve the aims and purposes of fair social cooperation over time. They are essential to preserve background justice, such as the fair value of political liberties and fair equality of opportunity, as well as to make it likely that economic and social inequalities contribute in an effective way to the general good, or more exactly, to the benefit of the least-advantaged members of society.
>
> -John Rawls (2001, p. 52)

Rawls's second principle of justice is divided into two parts. Broadly speaking, the principle on the whole addresses what is typically referred to as *distributive justice*. Given that the basic liberties of equal citizenship are secured, what are the further necessary design parameters for a basic structure that will assure the continued existence of a society of free and equal citizens? It is not enough that everyone is assured certain basic freedoms; everyone must also be assured of the conditions that enable the exercise of those freedoms. Just because everyone has the freedom to run for political office does not mean that society is structured so that this freedom is worth the same to everyone (e.g., in the US, only a select subset of citizens can *in fact* run for office because there are no checks on the expense of unregulated privately funded campaigns). Beyond securing an equally adequate scheme of freedoms for all, a just basic structure must also secure the equal value of those freedoms. All must be in a position to *exercise* their freedoms, not merely to *have* them (Habermas, 1996). This requires designing social

institutions that will assure a fair distribution of those resources that are essential for exercising the liberties of equal citizenship.

It is important to understand that for Rawls, distributive justice is not about how to divide up a pot of money or some collection of material goods resulting from social cooperation. He offers instead the idea that there are certain *primary goods* that must be fairly distributed in order to assure everyone the status of free and equal citizenship. A list of primary goods evolved during various iterations of Rawls's theory, including among other things: money, education, health care, and the social bases of self-respect. However, enumerating such a list is not the best way to understand primary goods and their fair distribution. Early on, Rawls offered a broader definition of what he had in mind (1971, p. 79): primary goods are those things that any reasonable person would desire, regardless of whatever else they desire. That is, primary goods are the "all-purpose means" of equal citizenship—what *anyone* would need to pursue *any* reasonable conception of the good as a member of a just society. Related to the philosophical ideal of moral agency discussed above, which framed the scheme of liberties everyone is entitled to, primary goods are those resources needed by each to allow for the continued equality of all.

Primary goods are thus the most appropriate basis of social comparisons for determining the positions of the best and worst off in society. Making social comparisons based on an index of Rawlsian primary goods takes into account more than just wealth, which is only one among many primary goods affecting the ability of individuals to exercise their freedoms. For example, consider a society in which the least well-off financially are given greater wealth in exchange for their right to vote and their freedom to choose an occupation (creating depoliticized but comfortably middle class citizen-employee-slaves, who would otherwise be that society's poorest). Such an arrangement, while raising their "standard of living," would not change the

status of this group as the least well-off according to an index of Rawlsian primary goods. Stripping a class of individuals of their political freedoms assures their unequal status as citizens and undermines the social bases of their self-respect, radically counteracting whatever freedoms might accrue from their improved financial situation (such as greater freedom as consumers). Recall the difference Rawls points out between the *liberalisms of happiness* (which might allow for such an arrangement) and the *liberalisms of freedom* (for which no improvements in "aggregate happiness" justify limiting the basic freedoms of equal citizenship). None of this is to say that poverty is not devastating and disempowering. In fact, Rawls does more than almost any contemporary political theorist to argue for major redistributions of wealth, including the need for a just society to establish a social minimum. The point is that we must be concerned with a broader set of primary goods—defined as the all-purpose means of equal citizenship—when considering questions of distributive justice.

This approach allows for more complex and meaningful questions about what constitutes a *fair* distribution of goods, transcending but including concerns about wealth. While the first principle of justice requires that a society's basic structure secures an equally adequate scheme of basic liberties for all, the second principle requires that this basic structure also gives appropriate shape to those inequalities that inevitably result from the activities of individuals as they exercise their basic freedoms. The goal is not create *perfect equality* where everyone has the same amount of everything (as in some utopian socialism), but rather to create *fair inequalities*, such that irrespective of where individuals are in the range between the best and least well-off, they understand themselves and their fellow citizens as participants in a fair system of cooperation—that they are each where they are as a result of institutions that all would agree to participate in. Rawls is concerned with specifying permissible ranges and types of social inequality,

specifically focused on securing each person's status as a free and equal citizen. Therefore the second principle can be read as suggesting that *equality of opportunity* is necessary for creating a society in which inequalities are fair. Moreover, because even fairly earned advantages tend to create unfair situations in the long run (due to the cumulative accrual of benefits to those with greater advantage), inequalities must be structured so that they always benefit the least well-off in order to secure continued fairness.

As with the first principle, these design parameters for just basic structures can be used to illuminate lessons from historical metrology about the institutionalization of measurement infrastructures. The near universal institutionalization of the metric system advanced the cause of securing everyone's right to objective measurement. However, just as the guarantee of basic freedoms specified by the first principle requires the second principle to assure the equal value of those freedoms, so the institutionalization of objective measurement requires further ethical oversight to assure its fair use. In particular, objective measurement can be (and has been) a means of oppression in situations where those subject to measurement are barred from opportunities to shape the measurement practices that affect them, often resulting in situations where the least well-off benefit the least from their right to objective measurement. In these situations it is often the very objectivity of the measure itself—and the scientific and political experts this objectivity requires—that systematically disempower people who, more often than not, had previously controlled their traditional measurement practices. The clearest examples of this can be found where the "metric system was spread on the end of a bayonet," as colonial powers reorganized and thus re-measured Africa, or as newly formed nation-states exercised metrological power over their own populations (Kula, 1986; Scott, 1998).

Early efforts toward state-organized agriculture in colonial Tanzania and the newly formed Soviet Union exemplify how objective measurement can be wielded in ways that create systematic injustices (Scott, 1998). In both cases, agricultural villages built using traditional measurement practices were dismantled and reassembled in terms of the new universal and objective metric system. This involved bringing all residences into a single "town center" and allotting the surrounding farmland as subdivisions of equally sized plots. State officials with modern surveying equipment could, for the first time in history, assure that each village was geometrically identical and that each farmer, no matter where they lived, had the same-sized plot of land. This was done in the name of both justice and efficiency. Governments understood themselves as overcoming the biases of traditional land allocation practices that reflected ancient ties of blood and place. They also sought to improve their administrative abilities by gaining precise knowledge of the size and number of farms, the kind of objective and reliable knowledge (of land thousands of miles away) necessary for modern taxation and state planning. These efforts failed, resulting in famines and the desertification of once fertile farmland. They also resulted in the creation of a new disenfranchised class, peoples who had been stripped of their power to administer their own lands and to determine the parameters according to which its value is measured.

Traditional measurement practices for the allocation of land in these communities had evolved over hundreds of years and rarely centered on considerations about equality of land size. They focused instead on properties of the land that were more relevant to the values of the community and the unique qualities of the land itself. For example, it was common to allocate land based on what was needed to sustain a family, which was not a specific area defined by metric units, but an area deemed appropriate based on a consideration of which crops could be

planted and the quality of the soil. In regions where highly nutritious crops could easily be grown, families received smaller plots; in areas with poor soil amenable to less nutritious crops families received larger plots. Similarly, it was common practice to create highly irregular plots depending on local geographical variations, such as the need to accommodate large boulders, steep hills, or seasonal floods. Farmers would be granted several discontinuous plots, or plots with boundaries that varied depending on seasonal weather patterns that made part of the land unworkable for periods of time every year. None of this is to suggest that traditional land allocations were always just. In fact, local variations often embodied the interests and biases of local leaders or the prejudices and superstitions of local cultures. The point is that redesigning these practices in terms of abstract, universal, objective units of measurement often resulted in even more profound injustices.

One infamous project involved the detailed planning of a large-scale agricultural community—from the size of the plots to the distance between the rows of planted crops (to accommodate newly standardized farming machinery). This planning was done in a hotel room thousands of miles away from the location of the community by administrators who had never stepped foot on the land (Porter 1995; Scott, 1998). Empowered by the assurance of accurate and objective measurement, they dispatched field agents to enact their plans and calculated predicted yields based on the average per-acre productivity associated with the new farm machinery, which they also sent out. Farmers protested the reorganization of their land and the new tools they were mandated to use, tools that required the standard measurement of each plot and row. These measures were essential for the efficient administration of the project from a distance, but useless on the ground, where, as any local farmer would have explained if asked, the difference between plots only miles apart were significant, despite their being the same size. Not only did the project

fail to produce the expected yields, it also created an underclass of disenfranchised farmers, newly dependent on the expertise of agricultural scientists, land surveyors, and machinists. By unilaterally changing the measures used to understand and manage the land, a basic structure of social life was put in place that systematically disempowered those most affected by it—farmers were stripped of their ability to use the measures that mattered to them.

Importantly, the allure of objectivity and its ideological relations to simplistic notions of distributive justice and efficiency created the impression among those implementing these initiatives that they were greatly improving upon existing conditions. Efficiency and justice were the ideals guiding the proliferation of measurement and standardization that came to dominate newly centralized agricultural production at the dawn of the twentieth century. Similar practices would spread through most major industrializing enterprises, including education, as is discussed at length in the following chapters. Justice was served through the efficient optimization of output from farms (or factories or schools), an aggregate good that justified overriding the autonomy of farmers (or workers or teachers) to assure food (or goods or skills) for the rapidly expanding urban populations. Scientifically promoting efficiency in industrial production was a massive international movement led by Frederick Taylor, which involved the extensive use of measurement practices and reorganizing of the lives of workers around these new objective measures. The momentous revolution inspired by the efficiency movement (Taylor was often compared to Marx, even by Stalin) involved an ethical argument about *the justice of efficiency*—that a "perfectly efficient system" is by definition just, because it wastes nothing and allows all to contribute to the whole. When it is held as the dominant institutional imperative, efficiency overrides individual autonomy in the name of administering justice for the sake of the whole and in the long run, using means that may or may not be just to those currently part of the system.

This complex relation between efficiency and justice, and the intensive role of measurement therein, will be a central theme in the analysis of educational measurement in coming chapters.

The idea that efficiency *is* justice can bring about exactly the form of injustice Rawls is concerned about, basic structures that unfairly limit individual freedoms in order to promote impersonal aggregate gains in happiness. It is unjust, according to Rawls, to compromise the basic freedoms of some in order to assure the free exercise of those freedoms by others (be they future generations or different social classes). Rawls frequently discusses the relation between efficiency and justice, and the temptation to simplify things by reducing justice to efficiency:

> There are many efficient arrangements of the basic structure.... The problem is to choose between them, to find a conception of justice that singles out one of these efficient distributions as also just. If we succeed in this, we shall have gone beyond mere efficiency yet in a way compatible with it. Now it is natural to try out the idea that... all efficient arrangements are [to be] declared equally just. But the suggestion seems... unreasonable [as a design parameter] for the basic structure..... The principle of efficiency cannot stand alone as a conception of justice. (1971, pp. 61-62)

The measurement practices accompanying "the cult of efficiency"—often involving scientifically objective instruments and universal standards—provide an impetus for the second and third principles of just institutionalized measurement. Simply put, in order to assure the fairness of a society's basic structures, all must have *a right to participate in determining what measures are relevant to them*, and moreover, *a right to benefit from measurement practices that*

involve them. This does not mean that individuals can simply reject measures that they don't like, such as those that cast their practices in a bad light. What it means is that the individual most affected by a measurement practice are entitled to have some say concerning the nature of how it is institutionalized, especially to assure that they benefit from it and are not merely subject to it while benefits accrue to those who unilaterally instituted it. Another way to put this is to say that all measurement practices must be acceptable from behind the Rawlsian "veil of ignorance"—e.g., could a measurement practice be reasonably be accepted by everyone involved if they were ignorant of their position with respect to that practice?

It is not enough to assure objectivity for all if only some get to decide what properties are to be objectively measured, while the rest are simply told what properties of the world now officially exist. This undermines equal opportunity by disempowering those whose metrological needs are left undefined—those who are thus rendered invisible to the administrative bureaucracies of the basic structure. Unjust modern bureaucracies are built and run using measurement practices that are objective, and often shown to improve efficiency, yet which are imposed upon the persons who must live and work according to them. They are imposed because they would not be used, or would be seriously altered, by those who depend upon them as a part of their practice. Their objectivity is undeniable and they thus satisfy the first principle of institutionalized measurement, yet they render the basic structure unjust nonetheless, because they are still used as an aspect of social practices that constrain the freedom of some to advance the freedoms of others.

The case of the unjust miller, discussed earlier in this chapter in conjunction with Rawls's first principle, taught the lesson of each person's right to objective measurement. This form of metrological injustice is often the result of sheer power and dishonesty, as objectivity is

overridden by *fiat*. The case of the *unjust bureaucracy,* discussed here in light of Rawls's second principle*,* teaches a lesson about the relevance and meaning of measures, beyond their mere objectivity—they shape the lives, opportunities, and self-respect of everyone using them.

In the coming chapters, this second principle of just institutionalized measurement will be re-articulated in the context of educational measurement, where it frames considerations of a test-based meritocracy built by a national testing agency and a campaign of national test-based accountability and privatized reform undertaken in the name of social justice. In these cases the tensions between justice and efficiency (including the ethos of *justice as efficiency*) are played out, revealing that the tools used by social justice reformers are at times themselves instruments of injustice.

Excursus on objectivity and the three principles of just institutionalized measurement

> There are several equally useful (often in different situations) notions of objectivity…. Some speak of objectivity as the ability to measure things precisely and accurately. Of course, critics might point out that precision and accuracy may be misplaced… Some understand objectivity as the avoidance of human subjectivity, its excision from a given situation…. Some see objectivity as something that emerges out of a community of practitioners… Some see objectivity as conformity with certain natural processes…as the length of the year is defined in reference to a natural phenomenon over which we have no control…. Yet another approach is "mechanical objectivity…the use of (usually) [automated] numerical techniques and procedures to reduce human judgment to

> such a degree that it is unnoticeable… hence the employment of mechanical objectivity serves as a barrier to unwanted criticism.
>
> -Lawrence Busch (2011, pp. 68-69).

The discussion so far has focused on the institutionalization of systems and instruments for measuring physical objects and has glossed over several complex issues in the field of *metrology*—the scientific study of the properties of measurement instruments. In particular, the concept of *objectivity* has not been treated adequately, nor have issues of *reference* and *representation*, the degree to which a measure refers to and represents what it claims to be about. However, because the focus of this work is on psychological measurement—and important distinctions between psychological and physical measurements systems are introduced in Chapter 3—it makes sense to focus any technical metrological discussions on *psychometrics*, the sub-branch of metrology that focuses on psychological measures. This excursus will clarify working definitions for important properties of psychological measures and explain how these interface with the three principles of institutionalized measurement discussed in this chapter.

Generally, discussions about objectivity (e.g., lack of bias, accuracy of score, context insensitivity), fall under the heading of *reliability* and are the related to the first principle, that is, the right to objective measurement. Whereas concerns about reference and representation fall under the heading of *validity*, and are related to all three principles. The broad distinction between *reliability* and *validity* is canonical in the field of psychometrics. But while there is agreement about this general differentiation of psychometric concerns, there are many different ways to unpack validity and reliability in more specific terms. Of particular use in this regard are the *Standards for Educational and Psychological Testing*, a document that has been recursively

revised for over three decades by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (1999). The *Standards* represent the considered judgments of literally hundreds of disciplinary experts about the kinds of quality control standards that ought to be in place for psychological measures of various types; they guide the discussion that follows.

Reliability and the virtue of objectivity

Being concerned about *reliability* entails looking into how well an instrument (e.g., a multiple choice test, a system of rubrics, etc.) functions as a measurement tool. Included among the various sub-categories of reliability are inter-rater reliability, internal consistency reliability, and test-retest reliability. This is by no means an exhaustive list, and definitions for all these terms and more can be found in any standard reference (e.g., Colman, 2001; AERA et al, 1999). Yet for our purposes here, suffice it to say that all forms of reliability concern issues of *measurement error*. Another way of saying this is that reliability deals with the properties that bear on the *objectivity* of a measure. Being concerned about reliability means looking into a measures' potential biases, its reasons for error-proneness, or inconsistencies in the results its yields due to contextual factors.

For example, *internal consistency reliability* is a quantitative index of how the items or levels of a measure function in relation to one another. Any psychological measure can be thought of as a system of categories used in the classification of performances. Insight into the internal consistency of a measure gives us a sense of how much noise—how much *measurement error*—surrounds each of its categories. Importantly, if the measurement error is too great—the categories are too fuzzy or noisy—then this means that the categories cannot be reliably

distinguished from one another, and this cuts down on the accuracy of the measure. For example, if the likelihood of an individual's performance being classified according to one of two adjacent categories is no greater than chance then these categories do not objectively discriminate differences in performance. That is, it is just as likely an individual will receive one score as the other, irrespective of the qualities of their performance that are the focus of the measure. Such a measure lacks the objectivity necessary for use in many contexts. Consider how inappropriate it would be if these two adjacent but non-distinct categories straddled the 'cut score' for admission decisions.

Inter-rater reliability also bears on issues of measurement error and objectivity. This form of reliability concerns the measurement error surrounding an instrument as it is put to use by different human raters. In other words, how much consensus is there between two or more raters when they award a particular performance a particular score? Low inter-rater reliability (i.e., when the same performance is likely to be given difference scores by different raters) means that that the results yielded by the measure are error prone, whereas high inter-rater reliability is an index of the opposite (i.e., the same performance is likely to be given the same score by different raters). Essentially, high levels of inter-rater reliability are needed in order to claim that a measure is objective; measures that lack high levels of inter-rater reliability lack objectivity.

It is worth briefly considering some of the various ways that assessments are scored to see the importance of inter-rater reliability as an aspect of objectivity. At one extreme are assessments that are scored automatically by machines, which removes the issue of inter-rater reliability from the equation entirely (assuming, that is, the machines are calibrated and maintained to function identically). In fact, this is one motive for advancing the computerized

scoring of wider ranges and types of performances: it produces a certain form of indisputable objectivity, referred to by Porter (1995) and discussed in later chapters as "mechanical objectivity"—i.e., because they are scored by machines, each performance is handled in *exactly* the same way, removing (almost) all chance of bias and human error in scoring (of course, this does not remove whatever bias there may be in built into the test question themselves). Then there are forms of test scoring that require human raters, but which are low-inference and quasi-mechanical, as when a teacher grades a test by hand that consists of simple short answers. Here there are issues of inter-rater agreement, but they mainly concern the probability of human error (e.g., losing track of which item is being graded and incorrectly marking a wrong answer as right, or vice versa). There are then a variety of relatively high-inference rubric-based systems, which is where most concerns about inter-rater reliability arise. These systems involve making complex discriminations about various aspects of performances, such as the quality of the writing, the strength of arguments, or the adequacy of one's conceptual knowledge and understanding. Importantly, these forms assessment strive for objectivity and often require that human raters undergo training so they are more likely to apply the system of categories consistently. Yet while these systems will never be as purely objective as a machine scored system, they can achieve very high levels of inter-rater reliability and thus be reasonably thought of as objective assessments. Finally, there are forms of assessment that involve non-repeatable unique performances, such as an essay written on a topic of a student's choosing. These kinds of assessments make much more limited claims to be objective in a technical psychometric sense, although they certainly make claims to be non-arbitrary. At the opposite end of the spectrum from the mechanical objectivity of machine scored assessments, these individualized one-off assessments also raise limited concerns about inter-rater reliability—the concern is not the

objective treatment of the student performance, but rather the appropriateness of the individualized feedback (although typically the idea is that another qualified reader would provide similar, although not identical, feedback).

Both of these forms of reliability (internal consistency and inter-rater agreement) are part of the conception of objectivity that is used throughout the rest of this work. As discussed below, certain aspects of validity also bear on objectivity, but it is worth discussing a few more points about reliability as an aspect of objectivity first. The first principle of just institutionalized measurement states that individuals have a right to objective measurement where it is possible and when it is preferable to existing alternatives. When the measure in question is psychological in nature, this means that *reliability must be considered as a social justice issue*. There are, of course, many situations in which the use of an objective measure is not preferable to existing alternatives (e.g., given the goals of graduate study, a doctoral thesis is not better replaced by a standardized test). Even so, in a wide range of educational situations the use of reliable objective assessments is both possible and preferable, from admissions tests to special education diagnostics.

In the case of admissions testing (or any form of testing that is used to classify and track individuals) the ethical implications of reliability are easily understood: the fair treatment of individuals requires that the test be objective. The categories established by the test should be truly distinct, otherwise scores could be erroneous or due to mere chance. Similarly, administration and scoring procedures should be appropriately standardized and unbiased, otherwise scores could be due to contextual factors or the biases of human raters. As discussed in Chapter 4, the institutionalization of test-based admissions and tracking systems was often justified in terms of the relation between objectivity and justice—echoes of the reformers who

understood the social justice boon that would accompany the spread of the metric system. However, reliability is not enough to assure a just testing infrastructure; concerns about the validity of the test are also paramount (e.g., Does it measure what it claims to? Does it measure what ought to be measured?).

In the case of special education diagnostics the need for reliability is equally important, although perhaps less obvious. These kinds of tests not only sort and classify students (justifying the availability of special resources; tracking students into special classes, etc.), they also inform an ostensibly scientific reinterpretation of a student's abilities and experience. Diagnoses from such tests suggests certain treatments and result in the application of (often emotionally-charged) labels to the student, which makes reliability ethically desirable because of the impacts these results may have on individuals' self-understandings, as well as how they are perceived and treated in educational institutions and beyond. Unlike the everyday construction of self-understandings and the ways in which students are understood by others in their day-to-day interactions, diagnostic tests give an authoritative 'scientific stamp of approval' to certain labels. The prospect of being mislabeled due to an unreliable diagnostic test is therefore of some ethical consequence: it will result in the wrong "treatment" and a misallocation of resources as well as create a false sense of who the student is and what they are capable of doing.

One final point about reliability must be raised before moving on to discuss validity. As mentioned earlier in this chapter, objectivity does not require quantitative representation. That is, highly reliable assessments can be entirely qualitative, as is the case with a well-built system of rubrics that yields high levels of inter-rater reliability. As discussed in Chapter 3, many contemporary bureaucratic contexts place a premium on the quantification of student performance, and this has impacted the way testing is both conceived and institutionalized. But

there is no need to limit objective testing to these demands for quantification. A wide range of highly objective non-quantitative options exist, although as will be discussed in later chapters, implementing these options would often require a substantial reallocation of resources and a re-conception of the purposes of testing in schools.

Validity and the need for meaning and relevance

Concerns about the *validity* of a measure are concerns about the legitimate inferences that can be made on the basis of the results it yields. These are concerns about the degree to which the measure represents what it claims to and thus performs the functions for which it was built. There are many different kinds of validity including but not limited to, construct validity, content validity, ecological validity, and predictive validity (Colman, 2001; AERA et al, 1999). Depending on the metric in question and the uses to which it is being put, different aspects of validity become relevant. Yet at their core, all types of validity address the question: *does the evidence support the conclusion that the instrument is measuring what it claims to be measuring?* It is important that the results yielded by a measure be valid—that we can be sure they are an index of what we think they are—so that we can make reasonable and responsible inferences based on them.

It is generally agreed that *construct validity* is the most important dimension of psychometric concern (Messick, 1980). Evaluating a metric in terms of its construct validity entails looking into various types of evidence concerning how the results yielded by the metric conform to expectations about the postulated construct or trait it is meant to be measuring. For example, if the instrument is intended to measure a relatively fixed aptitude—as the SAT is intended to do—then the results it yields should display specific patterns, e.g., they should not

vary as result of narrowly focused test preparation (which, of course, the SAT does, undermining the validity of the claim that it is a measure of aptitude: Lemann, 1999). Likewise, if a measure claims to be an index of future behavior (an aspect of construct validity called *predicative validity*) then the results should bear a certain relation to certain aspects of future behavior. For example, if the measure is taken as an index of future success or failure in college—as the SAT is—then the results should bear a strong correlation with academic performance in college (which the SAT does not: *ibid*). An instrument that lacks construct validity gives us reason to doubt that the inferences made based on its results are sound because there is little evidence that it actually measures what it claims to measure. An instrument with good construct validity, on the other hand, allows us to make well-justified inferences, guided by evidence-based beliefs that it is, in fact, measuring what we claims to be measuring.

On the one hand, construct validity bears on objectivity because claims to objectivity involve claims to be representing a specific quality—objectivity is always *about* something. Standardized administration and scoring do not necessarily yield objectivity in cases where, for example, a reliable instrument is used under the pretext it is measuring something it is not (as is the case with the SAT when it is taken as an aptitude test that is predictive of college success). There may be high-levels of inter-rater reliability and clear evidence of internal consistency, and yet the instrument is simply not being used correctly. To use an analogy with physical measurement: no matter how accurate our instrument for measuring the size of someone's foot, this instrument will never be an objective way to determine (or predict) an individual's height. Importantly, claims that foot size is a good *proxy* for height require evidence that this is the case, which in turn necessitates some independent way of measuring height. These kinds of claims—that measures of certain qualities serve as accurate but indirect proxies for a quality of interest—

are common in educational measurement and must be considered critically. In short, if there is no independent measure for the quality of interest, how can we ever know that the proposed proxy serves its function? This issue will be returned to in later chapters. In any case, because of its relation with objectivity *validity must be considered as a social justice issue*.

On the other hand, independent from its bearing on objectivity, validity is also directly related to the other two principles of just institutionalized measurement, which are concerned with the relevance and benefits of measurement infrastructures. Concerns about validity go beyond questions about the whether a measure represents what it claims to; they also raise questions about whether a measure does what it *ought* to do given the situations in which it is used. That is, given the context of a measure's institutionalized use, questions must be raised about the degree to which the measure plays an appropriate role in accomplishing the goals for which it was institutionalized in the first place. For example, an assessment may be good for measuring student reading ability—having high reliability and construct validity—and yet be used as an index of teacher quality. Or an assessment of vocabulary—good as far as it goes—is institutionalized as part of a program for tracking students into reading groups. In these cases, it is not that the assessment does not measure what it claims to, the issue is that it is used in questionable ways. We must ask: are the results relevant given what they are being used for? Does the use of the measure bring the benefits that are intended? Measures that are valid and reliable for some purposes are irrelevant and harmful when used for others. This further implicates validity as an aspect of social justice: the right to relevant and beneficial measurement entails asking a broad set of questions about the validity of measurement practices.

The goal of this excursus has been to clarify what objectivity means in the context of the discussions that follow. Essentially, objectivity combines reliability and validity: a measure is objective when it represents what it is claimed to represent (validity), and when it does so in an accurate, consistent, and unbiased manner (reliability). Measures that are characterized by this kind of objectivity are a prerequisite for social justice in a wide range of contexts.

2: Social justice and education

> [In a just society] education itself has been regulated by the principles of right and justice to which all would consent.... Thus no one's convictions are the result of coercive indoctrination. Instruction is throughout as reasoned as the development of understanding permits, just as the natural duty of mutual respect requires.... Moral education is education for autonomy.
>
> -John Rawls (1971, pp. 451-452)

> Education is the fundamental method of progress and reform.... The community's duty to education is, therefore, its paramount duty.... I believe every teacher should realize the dignity of his calling; that he is a social servant set apart for the maintenance of proper social order and the securing of the right social growth. I believe that in this way the teacher always is the prophet of the true God and the usherer in of the true kingdom of God.
>
> -John Dewey (1887, p. 95)

Just as measurement has been a social justice concern since the beginning of civilization, so has education. The first great writings on justice, from Plato to Confucius, focus on the education of the individual and the functions of schooling as an aspect of government. For over two centuries, the US public school system has been at the center of debates about social justice, with reformers, politicians, and teachers all at times understanding themselves as administering

justice, while at other times being accused of perpetuating injustices. Education has been understood as one of the royal roads to justice, by the philosophers mentioned above, as well as Rousseau, Kant, and Dewey. Schools have been the focus of billions of dollars of philanthropic money, millions of hours of scientific research, and thousands of policies, laws and regulations. Despite this attention, and the hopes pegged on the power of education, schools are still broadly understood as being on both sides of justice—as being both a source of justice and a source of injustice. Questions about schooling, education, and justice are still vexed; yet today large-scale testing-intensive reform efforts are being undertaken in the name of social justice. This chapter draws out essential lessons from Rawls's theory of justice for the philosophy of education. In Chapter 3 these lessons are integrated with the principles of just institutional measurement outlined in Chapter 1, resulting in a broadly applicable *theory of just educational measurement*.

The discourse about the relationship between social justice and education is ancient and complex. This is in part because of the "anthropologically deep-seated" nature of education (Habermas, 1971; 1984). All societies have had (and always will have) some form of education, and all people are educated in some way by someone. There is no getting around education, and in complex societies there is no getting around schooling, because it plays a necessary function in the perpetuation and re-creation of all complex social systems. It is no surprise then that Rawls considers educational systems as *basic structures*—understood as being subject to evaluation in terms of social justice, and as playing a critical role in the formation of a society of free and equal citizens. As discussed in the previous chapter, basic structures (unlike freely joined associations) are institutions joined by virtue of entering into society at all. They make up the unavoidable societal architecture according to which everyone must arrange their lives; the rules of the cooperative game we play by virtue of being alive. For this reason the basic structures

must be built carefully, to ensure that all who are implicated in them are treated fairly. Design principles for the creation of just basic structures are the focus of Rawls's *Theory of Justice*. Their application in the philosophy of education has been surprisingly sparse, considering the centrality of education as a contemporary social concern and the power of Rawls's theory (although see: Gutmann, 1987). What follows is a brief overview of a Rawls-inspired philosophy of education, focusing on only three essential issues that provide what is needed for the discussion of educational measurement to follow.

Rawls does not offer an extended discussion of education in any of his works, but he does offer a smattering of brief and complex statements about it. Combining these with the reasonable implications of his broader theory leads to a philosophy of education that is centered around *the role of schooling in a well-ordered society*. The idea of a well-ordered society, as discussed in the introduction, is one of Rawls's central constructs. Like the idealized person as citizen and the idealized decision procedure of the original position, it is intended to clarify complex intuitions and ideas about social justice. All of these "representational devices" allow us to see things more clearly because they simplify things, serving as a model or idealization of reality—similar to the role played by models in physical sciences such as chemistry and physics. Thinking about the nature of a well-ordered society is a way to clarify theories of justice by modeling how they work under ideal conditions. Rawls calls this "realistic utopian thinking," because it asks us to make a disciplined use of our imagination, following a structured process by which to test social ideas.

For Rawls, a well-ordered society is one that perfectly instantiates some conception of justice. Its citizens agree to a conception of justice that will govern them and all are aware of this consensus. They also believe, and are correct in believing, that the main institutions of their

society are governed according to the principles of justice they publicly agree to, which leads citizens to have a sense of justice they are motivated to follow. This is a "regulative ideal" found nowhere in reality—a society of free people reflectively self-administering justice. What theories of justice could be agreed upon by all and sustain such a society indefinitely? Rawls argues that theories maintaining the equivalence of efficiency and justice could not, nor could the utilitarian views associated with the liberalisms of happiness. In essence, these theories require too much of people—typically too much self-sacrifice on the part of the least advantaged—to be used successfully in the design of a well-ordered society. In the long run, everyone cannot agree to the fairness of their lot, as the suppression of freedoms for some yield gains in efficiency, aggregate happiness, and greater freedoms for others. Designing basic structures so as to direct the fruits of social cooperation toward *some end other than justice* (be it economic growth, war, or the flourishing of the arts) will inevitably leave some people treated as *mere means*. Those who have been treated this way cannot be expected to reasonably consent to unfair terms of social cooperation, and a well-ordered society requires the consent of all. A theory of justice capable of sustaining a well-ordered society must treat everyone as an end-in-themselves, and it must be clear to all that this is the case. Otherwise public consensus will deteriorate, as reflective citizens (likely from the least well-off classes) come to realize that their society's basic structures are systematically and disproportionately disempowering some to the benefit of others.

The principles of justice offered by Rawls serve as design principles for a well-ordered society. They involve the creation of a system of equal freedoms, supplemented by equality of opportunity and the difference principle, aiming for the institutionalization of basic structures that assure each individual (regardless of their function or use to society) a self-understanding as a free person, equal to all others. Rawls claims that a well-ordered society could exist in

perpetuity if so designed, as a stably self-replicating social system, evolving toward ever greater and more complex forms of justice.

When Rawls reflects on education and schooling, it is often in the context of thinking through what education would look like in this "realistic utopia" of the well-ordered society. Considering education in this way clarifies an ideal of what education ought to be by drawing attention to the function of educational systems in the perpetuation of social justice. Below I focus on three central issues in extrapolating from what Rawls provides toward a more comprehensive philosophy of education: 1) school systems as institutions that provide for *educational primary goods*; 2) the role of schools as part of a basic structure enabling the establishment of fair equality of opportunity; and 3) the Aristotelian principle (which posits a universal human preference for exercising complex skills in non-alienated work) as it relates to human development and self-respect. These three components are those needed to provide context for the discussion of educational measurement to follow. They do not represent the full scope of a Rawlsian philosophy of education, the total articulation of which would be a much larger task, including at least considerations of relevant topics Rawls broached, such as justice between generations, the structure of the moral personality, and the nature of moral motivations.

Schools, justice, and the nature of educational primary goods

> Democratic society is peculiarly dependent for its maintenance upon a course of study [that is] broadly human. Democracy cannot flourish where the chief influences in selecting the subject matter of instruction are utilitarian ends narrowly conceived for the masses, and, for higher education of the few, the

> traditions of a specialized cultivated class. The notion that the "essentials" of elementary education are the three R's mechanically treated, is based upon ignorance of the essentials needed for realization of democratic ideals.
>
> -John Dewey (1916, p. 192)

In the broadest sense, the principles of justice themselves are the central educative force in a well-ordered society. This is because they are the source of the norms that shape society's most important institutions. Like Dewey, Rawls argues that all the institutions in a society are educative in the sense that they shape the knowledge, skills, and dispositions of individuals, often in profound ways. As discussed in Chapter 1, this is one of the main reasons to be careful in the design of a society's basic structures—they create people. But in a well-ordered society, the principles take on a wider role in their function as the focus of public agreement; the principles are educative in an even more explicit way. As Rawls explains:

> [In a well-ordered society] a political conception [e.g., of justice] assumes a wider role as a part of public culture. Not only are its first principles embodied in political and social institutions.... [The public culture] also contains a conception of citizens as free and equal. In this way citizens are made aware of and educated to this conception. They are presented with a way of regarding themselves that otherwise they would most likely never be able to entertain. To realize the full publicity condition is to realize a social world within which the ideal of citizenship can be learned and may elicit an effective desire to be that kind of person. (1996, p. 71)

Of course, there are institutions that are reflectively educative, that are built and perpetuated for the sake of education. The family and the school are the most important of these (libraries, museums, and some television programs are examples of others). Rawls understands both the school and the family as essential in the reproduction of society through the transmission of culture, knowledge, and skills. He suggests a kind of division of labor between the family and the school in the education of the future generations. In considering the role of education in a well-ordered society, Rawls demonstrates that his theory of justice involves an important distinction between the family and the school as educational institutions (Rawls, 1971; 2001). The publicly supported school system is an institution of the government designed to guarantee certain essential primary goods to all citizens. As discussed below, this sets it apart from other educational institutions, such as private colleges and universities, religious schools, and some entities like charter schools, which limit admissions and prioritize goods of a much wider variety.

The family as an institution is distinct from both private and public schools, while interfacing with both in complex ways as all three institutions co-evolve. The family in a well-ordered society is responsible for providing individuals with the meanings that give value to their lives, as well as other things, such as love, support, and companionship, all of which are essential to human development. According to Rawls, it is not within the state's right or capabilities to provide such things for it citizens, even though they are essential for their development, and thus essential for the continuation of the state. Rawls is at pains in several places to demonstrate that while families are subject to the principles of justice, the vision of the good life guiding individual families is not to be made subordinate to an overarching vision of the good life offered

by the society at large, especially the government. Families educate individuals in ways that should not and cannot be replicated by other social institutions, which is one of the reasons why Rawls argues for economic fairness with respect to domestic labor and a social minimum to avoid the damage done to families by poverty:

> The principles of justice apply to the family [because] the family is part of the basic structure, the reason being that one of its essential roles is to establish the orderly production and reproduction of society and of its culture from one generation to the next.... Accepting this, essential to the role of the family is the arrangement in reasonable and effective ways of the raising and caring for children, ensuring their moral development and education into the wider culture.... No particular form of the family (monogamous, heterosexual, or otherwise) is so far required by a political conception of justice so long as it is arranged to fulfill these tasks effectively and does not run afoul of other political values.
> (Rawls, 2001, p. 163)

Anthropological evidence suggests that the extended immaturity of the human organism demanded the participation of the father and the creation of the family unit in order to enable the raising of children, making the family the original institution founded for the sake of education. The family survived as the predominant source of education for the vast majority of human existence, with only sporadic (if interesting) attempts at universal schooling in particularly ambitious places, such as ancient Sparta. Universal schooling gained ascendancy only relatively recently, as the family receded as a center of production and the vast majority of household

heads were brought into the wage-labor system (Bowles & Gintis, 1976). The modern nation-state ushered in a new era in human development with the advent of mass schooling. In post-modern societies it is hard to imagine a time before schooling, let alone alternatives to large school systems in providing for the maintenance and reproduction of our complex economic and political systems.

In a well-ordered society, the publicly supported school system is intended to supplement what is provided by the family, in order to assure the equitable distribution of *educational primary goods*—the basic opportunities for human development guaranteed to all. Importantly, although Rawls argues for public funding and support for schools, he does not argue explicitly for a public school system *per se*—e.g., a single state run bureaucracy—and remains open to the idea of a publicly funded system of private schools (Freedman, 2007). But Rawls's lack of clear commitment to a state-run public school system is not an argument in favor of the *privatization* of schooling. He argues for a *system* of schools—not a marketplace of schools—coordinated and designed to ensure the just profusion of educational goods and opportunities. As with his views on health care, Rawls argues that providing for primary goods is a social responsibility that cannot be met using market mechanisms as the sole determinate of transactions and allotments. So, if it is a system of private schools that is to sustain a well-ordered society, it must be designed and regulated to ensure that its schools serve the public conception of justice, and thus cohere and contribute to the background justice of the broader basic structure. A system of private schools that remained economically stratified from generation to generation, for example, would segregate the socialization of social classes, and thus undermine the self-understandings and dispositions of free and equal citizenship—such a school system would be unable to serve as part of the basic structure in a well-ordered society (Rawls, 1971).

In a well-ordered society, schools must provide for the development of individuals who have all the basic capabilities needed to participate in society as free and equal citizens. Education is positioned alongside other basic rights and entitlements to primary goods, such as freedom of movement, freedom of conscience, right to due process, and entitlements to adequate food and shelter. There is an "amount" of education owed to everyone, which Rawls specifies as the amount prerequisite for social participation and for securing the equal value of political liberties. As Rawls explains, the constitution of a well-ordered society is "required to assure that the basic needs of all citizens can be met so that they can take part in political and social life.... Below a certain level of material and social well being and of training and education, people simply cannot take part in society as citizens, much less as equal citizens" (1996, p. 166).

Two responsibilities stand out as the function of schools in a well-ordered society. First, for a well-ordered society to reproduce itself, everyone must be provided with the education needed to give (or withhold) reasoned consent to the institutions and principles of justice that govern their society, and thus to participate in political life. Schools must provide what is needed for the development of citizens who "regard themselves as self-authenticating sources of valid claims. That is, they regard themselves as being entitled to make claims on their institutions so as to advance their conceptions of the good (provided these conceptions fall within the range permitted by the public conception of justice)" (Rawls, 1971, p. 23). Related to this civic and moral focus, schools must also provide the basic skills needed to pursue a reasonable self-chosen conception of the good life. Schools in a well-ordered society are to equip individuals with the "all-purpose means" needed to pursue as wide a range of values and visions of the good life as possible, given the constraints of what justice requires and of historical circumstances. This

second responsibility of schooling in a well-ordered society is discussed following a clarification of the first.

The first responsibility requires making schools into places where each student comes to understand the principles that govern the political life of their society, including how these inform the basics of political process, and their rights and responsibilities as citizens. Rawls argues for a school system built to provide a certain kind of civic and moral education—an education designed to perpetuate the continuation of society as a cooperative enterprise between free and equal citizens. This requires a general approach to curriculum and to the social relations and organization of the school, as both are to be built in light of a "political conception" of the student. "Society's concern with their [e.g., the students'] education lies in their role as future citizens, and so in such essential things as their acquiring the capacity to understand the public culture and to participate in its institutions" (Rawls, 1996, p. 200). Rawls further specifies the broad contours of the curriculum:

> [A well-ordered society] will ask that children's education include such things as knowledge of their constitutional and civic rights, so that, for example, they know that liberty of conscience exists in their society and that apostasy is not a legal crime, all this to ensure that their continued religious membership when they come of age is not based simply on ignorance of their basic rights.... Their education should also prepare them to be fully cooperating members of society and enable them to be self-supporting; it should also encourage the political virtues so that they want to honor fair terms of social cooperation in their relations with the rest of society. (2001, p. 156)

Moreover, the delivery of this curriculum must be taught so as to embody the relations of mutual respect that characterize the broader political culture. The principles of justice themselves "regulate the practices of moral instruction in a well-ordered society. Thus moral education is education for autonomy" (Rawls, 1971, p. 452). This is a notion raised several times in passing by Rawls (1971, pp. 183 & 220; 2001, p. 166). The broad idea is that a just society requires a certain approach to pedagogy, one that considers the student-teacher relationship and the policies and practices of the school itself as governed by the same principles that govern the society as a whole. Echoing Dewey, who argued that the school should be a "microcosm" of the larger society, Rawls argues that schools should be the places where children first come to know what it means to participate in an institution built according to the principles of justice that govern their society.

Designing schools in this way is necessary in order to assure that students do not adopt a view of justice that has simply been inculcated in them as a result of power differentials in the social relations of the school (e.g., through indoctrination or coercion). Students must come to freely understand—to construct for themselves in their own terms—the validity of the principles of justice, how they inform the basic structures of their society, and how their own rights and responsibilities as citizens are derived from them. Structuring the school as a 'microcosm' of a just society is also necessary because of the inevitable impact of institutional life on the development of character. If the majority of every citizen's young life is spent in an institution that does not reflect the political ideals and processes of the broader society, how can they be expected to develop the dispositions and self-understandings that prepare them to be participants in that society as adults? One of the great contradictions in American schooling has always been

the non-democratic structure of the schools, which have nevertheless been intended to prepare citizens for participation in democracy (a point returned to in Chapters 4-6). Rawls does not explore the implications and details of his curricular and pedagogical reflections, let alone apply them to the problems of modern and postmodern schooling. But the implications of these ideas are the focus of later discussions, where standardized testing is shown to be one of the main causes of radically truncated curricula and distorted and unjust social relations in schools.

The second responsibility of public schools in a well-ordered society—providing for the basic skills needed to pursue a self-chosen conception of the good life—is more subtle and complex than the first. As discussed in the previous chapter, Rawls views considerations about justice as being different from considerations about what is valuable or good. Beyond the requirements set by justice (e.g. one's pursuit of a good life cannot be predicated on limiting the goods pursued by others), citizens are free to give meaning, direction, and value to their lives in whatever ways they see fit. Schools share a large part of the responsibility for making this freedom a reality. By some readings this suggests that schools be designed as all-purpose opportunity providers, which is an impossible task. In fact, as Rawls explains, schools cannot provide the means for everyone to pursue anything and everything; they are only to provide the means to the further education required to advance a reasonable conception of the good life. Typically, other educational institutions are involved in an individual's self-chosen pursuit of a meaningful life, but schools are responsible for providing a way into the exercise of this basic freedom. This entails that schools be designed accordingly. They must be fair to competing values and worldviews by not unduly privileging one over another, while also providing a wide enough range of skills and a wide enough exposure to cultural differences that students

understand themselves as freely making reflective decisions about the value and direction of their own lives.

Generally put, in a well-ordered society, no conception of the good compatible with justice is to be made inaccessible through schooling, while no specific good compatible with justice is to be favored. That is, justice requires that a society's basic structures be fair to all reasonable values and ideals of the good life (Rawls, 1999). However, this does not mean all possible worldviews and values are present in society or open to students. "The full range of values is too extensive to fit into any one social world" (Rawls, 2001, p. 154). As Rawls explains, picking up a theme that is central to the work of Isaiah Berlin, "there is no social world without loss":

> No society can include within itself all ways of life. We may indeed lament the limited space, as it were, of social worlds, and of ours in particular; and we may regret some of the inevitable effects of our culture and social structure.... [There is] no social world that does not exclude some ways of life that realize in special ways certain fundamental values. But these inevitable exclusions are not to be mistaken for arbitrary bias or for injustice.... That there is no social world without loss is rooted in the nature of values and the world. Much human tragedy reflects that. A just liberal society may have far more space than other social worlds but it cannot be without loss. (Rawls, 2001, pp. 154-155)

Likewise, not all of the values and visions of the good life made accessible through schooling can be realized by an individual. Rawls considers this an unavoidable fact of social

life, and argues that the basic structures of society must thus be designed to coordinate the inter-animation of individuals who build their lives around diverse social enterprises and forms of life. In several places he offers the ideal of a harmonious "social union of social unions" as a way of characterizing the higher-order unity playing out through the coordination of diversity enabled by basic structures that are just. This provides a way of thinking about the goal of an educational system that is not built to promote one vision of the good, but rather to enable the flourishing of individuals who pursue a diversity of goods within a framework of just institutions that privileges none:

> In a fully just society persons seek their good in ways peculiar to themselves, and they rely upon their associates to do things they could not have done, as well as things they might have done but did not. It is tempting to suppose that everyone might fully realize his powers and that some at least can become complete exemplars of humanity. But this is not possible. It is a feature of human sociability that we are by ourselves but parts of what we might be. We must look to others to attain the excellences that we must leave aside, or lack altogether.... The good attained from the common culture far exceeds our work in the sense that we cease to be mere fragments: that part of ourselves that we directly realize is joined to a wider and just arrangement the aim of which we affirm. The division of labor is overcome not by each becoming complete in himself, but by willing and meaningful work within a just social union of social unions in which all can freely participate as they so incline. (Rawls, 1971, p. 464)

An educational system that can prepare individuals to reflectively navigate this kind of openness and plurality of goods requires the consideration of a wide variety of school practices (from tracking, grade, and ranking to curriculum design and pedagogy) in terms of how they shape the life-prospects and worldviews of students. Again, these are details that Rawls neglects, but which will occupy us in the discussion of educational measurement below, where it is shown how efficiency-oriented testing practices can unjustly limit opportunities for human development, ultimately allowing for access to a very limited range of life prospects—leaving students with unjustly truncated and homogenized conceptions of the forms of life that are possible and preferable for them to pursue.

Schooling and fair equality of opportunity

> Democracy is more than a form of government; it is primarily a mode of associated living, of conjoint communicated experience.... The widening of the area of shared concerns, and the liberation of a greater diversity of personal capacities which characterizes a democracy... [are] a matter of deliberate effort to sustain and extend. Obviously a society to which stratification into separate classes would be fatal, must see to it that intellectual opportunities are accessible to all on equitable and easy terms.
>
> -John Dewey (1916, p. 88)

That schools can both enable or bar access to different opportunities in life relates directly to the next major function of educational systems in a well-ordered society: to help

assure that there is fair equality of opportunity. The reproduction of a well-ordered society (like any other society) requires that all positions and careers be cyclically re-populated, as older generations pass on the responsibilities of maintaining society. Justice demands that the social mechanisms used to fill these positions are acceptable to everyone, which is why the reproduction of society through aristocratic and nepotistic means undermines the functioning of a well-ordered society. Tax codes affecting inheritance, political processes surrounding elections and appointments, as well as economic inequities, all play a role in the allotment of opportunities. But schools play a unique role in the reproduction of society. Because they administer what society owes each citizen in terms of educational primary goods, institutions serving *not* as a selective mechanism for putting individuals in their place, but rather to ensure that no one is kept from a place suited to them due to a lack of basic skills. Schools must assure equal access to opportunities to develop those skills needed to take a self-chosen place in society as a free and equal citizen—these are the skills that reliably result from access to educational primary goods. Throughout Rawls's work the principle of fair equality of opportunity includes education as a part of its definition: "As earlier defined, fair equality of opportunity means a certain set of institutions that assures similar chances of education and culture for persons similarly motivated and keeps positions and offices open to all on the basis of qualities and efforts related to the relevant duties and tasks" (Rawls, 1971, p. 245).

A version of this idea is, of course, the cornerstone of progressive educational reform and the lynch pin of arguments coming from those who imbue education with the power to transform all social inequities. That education can create a level playing field is one of the great refrains in the history of American education. Rawls maintains a complex version of this belief about education, but it is one that does not shoulder the educational system with accomplishing

democratic equity singlehandedly. Instead, he positions the school system in a network of other institutions, including political, economic, and social service organizations, which *together* provide a context of fair equality of opportunity that fosters the development of individuals who think and act as free and equal citizens. The school system plays only one function in establishing the overall background justice of equal opportunities provided by the basic structure.

The school system, in fact, requires that other institutions in the basic structure are arranged so that it can perform its function in fostering equal opportunity. Related to equality of opportunity and the role of education in the reproduction of society, Rawls discusses choice of occupation as a basic right, along with the right to non-alienated labor (which is discussed below), both of which require that the educational system function as part of a broader basic structure that equalizes economic and social inequity. There is a codependence between the institutions of the basic structure; for example, educational institutions of certain types are symbiotically related to political institutions of certain types, as the institutions of law require the existence of law schools. For a school system to serve justice it must be put in a position to do so by the other institutions of the basic structures that surround it. In particular, just schools require just economic and political establishments, which surround them and into which they send students upon graduation. If opportunities for non-alienated labor are scarce and political participation is limited to viewership and voting, then schools that promote non-alienated modes of learning and discourse-based school governance will not have a social ecosystem in which to thrive. Schools cannot overcome or counteract the broader structural injustices that affect the lives of students, teachers, and administrators.

Despite these clear relations between schools and the rest of the institutions of the basic structure, it is nevertheless a common expectation that the educational system can serve as the

great *panacea* for social inequality. Rawls argues that a *system of institutions* must be arranged so that the basic structure provides an adequate degree of background justice; educational institutions must be nested in a broader network of institutions that administers justice by design.

> The principles of justice apply to the basic structure and regulate how its major institutions are combined into one scheme. [The] idea of justice as fairness is to use the notion of pure procedural justice... The social system is to be designed so that the resulting distribution is just however things turn out.... [This entails that] the basic structure is regulated by a just constitution that secures the liberties of equal citizenship. Liberty of conscience and freedom of thought are taken for granted, and the fair value of political liberties is maintained.... [There must be] fair (as opposed to formal) equality of opportunity. This means that... the government tries to insure equal chances of education and culture for persons similarly endowed and motivated either by subsidizing private schools or by establishing a public school system. It also enforces and underwrites equality of opportunity in economic activities and in the free choice of occupation... policing firms and preventing monopolistic restrictions and barriers to the more desirable positions. Finally, the government guarantees a social minimum either by family allowances and special payments for sickness and unemployment, or more systematically.... (Rawls, 1971, p. 243)

The ideal of a social system designed to administer a kind of *pure procedural justice* is at the center of Rawls's thinking about the role of schools in securing fair equality of opportunity.

Pure procedural justice is one of Rawls's meta-ethical tropes—which stands alongside perfect and imperfect forms of procedural justice—as part of an argument that justice can be administered without a view to an ideal *outcome* but rather by adherence to an ideal *process*. If the rules are set up to be fair and they are adhered to, then the outcome is fair regardless. A social system built according to the principles of justice is one designed to administer pure procedural justice. Such a basic structure does many things, including legitimating the entitlements of individuals to whatever they are able to achieve within its structure of allowances and opportunities. A shared public conception that the social structure administers pure procedural justice brings with it the belief that each person's social position is in some way justified, or fair—that whatever grievances of fate, illness, and calamity befall individuals are not due to the impact of social institutions.

The school system is one of the central institutions that provides explicit reasons and experiences that justify (or create a public sense of fairness about) those inequities in the distribution of power and rewards that inevitably result as each new generation repopulates and transforms the social world. In a well-ordered society, the basic structure is arranged so that schools can function as institutions that equalize opportunity by uniformly providing every person the educational primary goods needed to develop into a free and equal member of society. Unique among the institutions of the basic structure, the school system in a well-ordered society serves as a kind of on-ramp into a network of institutions oriented toward a purely procedural allocation of rewards. The schools are the first places where children are expected to participate in an institution that places them in complex cooperative relations with others. It involves participation in a nuanced system of mutual-expectations, where the benefits of success include nearly immeasurable expansions in possible future rewards. Individuals take an immediate

interest in who does well and why, in how their participation is received by the group, and in how they are benefiting and expanding personal freedom as a result of their schooling. It is in their experiences as students that citizens first encounter the impact of socially-sanctioned institutional norms on the shape and feeling of their day-to-day lives, their self-esteem, and their ideals and hopes about their future.

Generally speaking, a just school system should result in a general public sense that those who reap the most benefits from education are those who are rightfully entitled to do so. The schools (again like a microcosm of the larger society) should be structured fairly—designed to administer a kind of pure procedural justice—so that every citizen can look back upon their education as a time when they were treated fairly by the society they were being prepared to enter. The citizen in a well-ordered society is able to claim responsibility for their abilities and the significant opportunities that they have (or have not) been afforded; they are made to feel neither that they were unduly privileged nor that they were the victim of institutional neglect or injustice. If pure procedural justice has been achieved, students will be able to look back at their educational biographies, and those of others, with a sense that everyone has gotten what they deserved.

Importantly, this does not necessarily lead to a *meritocracy*. Rawls uses this word pejoratively to refer to a society in which the "social order follows the principle of careers open to talents and uses equality of opportunity as a way of releasing men's energies in the pursuit of economic prosperity and political dominion" (1971, p. 91). As discussed in Chapters 4-6, there has been an unfortunate and often contradictory alliance between progressive education reform, standardized testing, and a broad belief in the justice and efficiency of educational meritocracy. Rawls is aware of this possible technocratic short-circuiting of schooling and argues that schools

should not serve as a sorting mechanism for allocating opportunities along narrow definitions of merit and achievement, definitions ultimately furnished by systems outside of the schools—primarily economic and political ones. Fair equality of opportunity does not mean "an equal opportunity to leave the less fortunate behind on the personal quest for influence and social position" (ibid.). Rawls continues arguing for a democratic (as opposed to economic) characterization of equal opportunity by invoking the difference principle, which requires that what inequalities do accrue to those more favored somehow come to benefit the prospects of the least well-off.

> The difference principle transforms the aims of society in fundamental respects. It transforms the aims of the basic structure so that the total scheme of institutions no longer emphasizes social efficiency and technocratic values.... Education should not be assessed solely in terms of economic efficiency and social welfare.... [This is clearest] when it is necessary to take into account the essential primary good of self-respect and the fact that a well-ordered society is a social union of social unions. It follows that a confident sense of their own self worth should be sought for the least favored, and this limits the forms of hierarchy and the degrees of inequality that justice permits. Resources for education are not to be allotted solely or necessarily mainly according to their return as estimated in productive trained abilities, but also according to their worth in enriching the personal and social life of citizens, including the least favored. (Rawls, 1971, p. 93)

Picking up on the theme of society as "a union of social unions," and on the perennial tension between efficiency and justice as institutional virtues, Rawls clarifies the implications of the second principle for the design of school systems. In just schools equality of opportunity means equal opportunity to assume a self-chosen place within a fair system of cooperation. It means an equal opportunity to develop a self-concept based on an accurate sense of self-worth, to develop an allegiance to justice, and to come to reflectively endorse and pursue an ideal of the good life. With the surrounding institutions in place, including economic opportunities for non-alienated labor, a social welfare minimum, and universal healthcare, schools no longer shoulder the burden of accomplishing democracy despite adverse economic and social conditions. The basic structure in a well-ordered society is one in which justice conditions the meaning of efficiency: it prioritizes taking into account the positions of the least well-off and redressing their misfortune by utilizing the energies of those in better positions.

According to this ideal, schooling fails if it is designed and administered as a zero-sum game; it fails even if this kind of competition efficiently provides the economy with exactly the workers it needs. Educational opportunity is not best thought of as an opportunity to "get ahead" (either individually or collectively), as if schools were distributing some kind of scarce resource, a view that only makes sense if what schools are distributing are job opportunities or other economic returns (and even in this case it is not clear that scarcity needs to be the predominant economic motivator). Rather, educational opportunity is best thought of as an opportunity to "know thy self" —to borrow again the well-worn Socratic imperative. This is a process enabled by (among other things) coming to hold progressively more complex understandings of the social and natural world, participating as a member of increasingly demanding cooperative enterprises, and thus learning about the unique role and value of each person, including one's

self. Schools must enable students to value themselves, irrespective of (but not ignorant of) what they are ostensibly "worth" to society. In a well-ordered society valuable humans are not understood as a scarce resource because human worth is not reduced to a simplistic meritocratic ranking. A fuller discussion of the social justice issues involved with meritocracies is presented in Chapter 5.

Schooling and self-actualization: the requirements of the Aristotelian Principle

> When it is said that education is development, everything depends on how development is conceived. Our net conclusion is that life is development, and that developing, growing, is life. Translated into its educational equivalents, that means (i) that the educational process has no end beyond itself; it is its own end; and that (ii) the educational process is one of continual reorganization, reconstructing, transforming....
>
> -Dewey (1916, p. 50)

The discussion above does not adequately deal with the problem that a school system can provide fair opportunities for access to what is an unfair range of life prospects. This raises the final set of issues in Rawls's philosophy of education, those surrounding the *Aristotelian principle*. This principle of human nature serves to animate what would otherwise be a static formal social structure for administering pure procedural justice. It posits a set of basic human motivations that internally direct the development of each individual in the social system toward increasingly complex and non-alienated forms of work. The idea is that when the basic structures

are designed correctly, there is a natural wellspring of human striving for meaningful learning and complex work. The Aristotelian principle is an important part of Rawls's philosophically motivated moral psychology, which has several other key components, such as a stage theory of moral socialization and a complex set of views on virtue.

Like many of the big ideas discussed so far, the Aristotelian principle is a kind of ideal or "useful fiction." Taking its name from the arguments offered by Aristotle (see Ross, 1921) on the relations between learning, enjoyment, and constructive activity in the *Nicomachean Ethics,* it is a "principle of motivation" that "expresses a psychological law governing changes in the pattern of our desires" (Rawls, 1971, p. 375). It is offered as what Habermas (1999, p. 14) would call a "reconstructive hypothesis" —a philosophical construct that stands in place of an "empirical theory with strong universalistic claims." In other words, it is part of a theory of human nature, one that conforms to the best of what is known in the relevant special disciplines (e.g., psychology, sociology, anthropology, etc.), while nevertheless transcending their limited claims in an integrative reconstruction fit for use in everyday life and political discourse (ibid.; Rawls, 1971, p. 375). "The Aristotelian principle runs as follows:

> Other things being equal, human beings enjoy the exercise of their realized capacities (their innate or trained abilities), and this enjoyment increases the more the capacity is realized, or the greater its complexity. The intuitive idea here is that human beings take more pleasure in doing something as they become more proficient at it, and of two activities they do equally well, they prefer the one calling on a larger repertoire of more intricate and subtle discriminations.... (Rawls, 1971, p. 376)

This is ascribed as a universal human tendency, not an invariant causal pattern. The tendency can be overridden, countervailed, and otherwise inhibited—it is also likely to be offset by contradictory psychological mechanisms, such as tendencies to stasis and the transformations of motivation that occur during aging and other forms of physical limitation. But Rawls only needs the principle to be "true enough." It is offered only as an orienting generalization about human development that can bring reasonable coherence to our considered judgments about how individuals behave when they are free to choose how to use their time. Although this is not the place to discuss findings from psychology, there is ample evidence to suggest that when relieved from the pressure to compete for survival through the institutionalization of, for example, a social welfare minimum or subsidies for universal college education, individuals do not choose to take the path of least resistance or to "loaf on the dole" (Lane, 1991).

The Aristotelian principle is a speculative guess at the riddle of what lies beyond the equally fictional economic motivation of pure "rational choice," which informs considerations by economists about how to design for social welfare. In a well-ordered society, the basic structures do not set up a system expecting to deal with humans who are disinterested actors that make decisions only to advance their own interests. Instead, individuals are taken to have an innate human desire to enlarge the complexity and range of their competence, and to thus expand the control they exercise over the activities in their lives, especially their work. This is another way in which Rawls brings in a minimal conception of the good, in order to design the basic structures to accommodate as much diversity as justice allows. The Aristotelian principle does not dictate what kinds of skills and competencies are to be developed—it does not outline some specific way to satisfy the drive to mastery—it says only that humans have this drive and so the

institutions in which they become people must be free enough to *not* thwart it, if not liberating enough to foster and catalyze it. It is another minimal (or formal-structural) characterization of what citizens strive for in a well-ordered society, aside from striving to administer justice.

Schools are the first places where individuals can feel this pull toward competence and the pleasure that comes from learning and increasing the complexity of one's skills and ideas. Beyond providing for a sense of justice, schools in a well-ordered society must also take account of this universal human impulse toward complexity, mastery, and non-alienated work. Schools are to be places where this motivation can receive positive reinforcement, be enlarged, and leveraged to bring each individual into the fullness of their autonomy and their unique role in the social system. This idea is explicitly opposed to the long-standing (if often disowned) belief that some people are simply not interested in developing complex skills, let alone exercising them in contexts over which they take responsibility. The Aristotelian principle contains a kind of reasonable optimism about the natural tendencies of human development, and posits an inborn desire to learn and work to the best of one's ability.

As mentioned above, the Aristotelian principle is part of Rawls's broader moral psychology—his theory of human nature—which takes from a tradition of psychological theorizing that views human development as an *epigenetic* process, a process of self-transcendence, where the self-system and its beliefs undergo qualitative reorganizations—co-evolving in dynamic relations with the social and physical world (Fischer & Bidell, 2006; Kohlberg, 1984; Piaget, 1932). This tradition is opposed to psychological models that make individuals out to be IQ-based information processors or perfectly rational economic self-optimizers. Instead, it represents the individual as an evolving, reflective, self-constructing person, who holds beliefs that change significantly during the course of life, and is driven by

diverse interests, including an interest in his or her own continued development. As a supplement to the Aristotelian principle, Rawls (1971, p. 403) offers a stage theory of moral socialization arguing for the broad value of what could best be described as *developmental complexity*. This theory puts forth that a certain type of human development needs to be accounted for in the design of the basic structure—and especially in educational systems—in order to assure that citizens are allowed the fullest expression of their humanity.

If all are to understand themselves as free and equal then society must secure the possibility of full self-actualization for everyone. If some citizens are made simple while others are made complex by the design of the basic structure itself, this undercuts their ability to see themselves as part of a system of fair cooperation. Of course, there are natural differences in people's abilities, propensities, and developmental potentials. Importantly, in most cases, given the limited range of diagnostic tools, there can only be a pretense of knowledge concerning such individual differences; so there is almost never a strong enough index of what someone is capable of to cut them off from developmental opportunities. This is a topic that will be returned to in the chapters to come, where testing is shown to have served just such a sorting and limited function in many contexts, some still current. The Aristotelian principle says that the basic structure is to facilitate (although it cannot guarantee or cause) self-actualization in the fully Maslovian sense (Maslow, 1968). A just education system seeks to promote the development of individuals with complex thought and emotion, who are masterful in an area recognized as valuable socially, and who understand themselves as self-authoring.

As with equality of opportunity, Rawls does not shoulder the educational system with the sole responsibility of providing for the Aristotelian principle's demands. The institutions of the economic system, where most people spend the majority of their lives, must allow for the kind of

non-alienated labor and lifelong learning opportunities that align with human tendencies toward continued development and the desire for increasing complexity and autonomy. The political and legal systems must allow for free self-expression, diversity of lifestyle, and convey an openness to cultural evolution, such that individuals can come to thrive in complex ways, perhaps exercising skills as a part of their self-actualization that have never been seen before nor will be seen again. Yet, obviously, schools do play an important (if not central) role in caring for the Aristotelian principle's demands. As with the other Rawlsian lessons for the philosophy of education explored in this chapter, he does not specify the implications of the Aristotelian principle for educational practice. Still, the principle makes clear demands on educational institutions and practices. These are explored in following chapters, where testing is shown to be one of the ways that the Aristotelian principle can be either catalyzed and leveraged, or short-circuited and squelched.

This chapter has focused on some of the essential insights Rawls's theory of justice has for the philosophy of education. Rawls offers three lessons about the nature of just educational institutions: (1) they provide for educational primary goods, (2) contribute to a system of institutions that secures equality of opportunity for all, and (3) make possible individual self-actualization. When this minimal philosophy of education is combined with the principles of just institutionalized measurement from Chapter 1, the basic outlines of a theory of just educational measurement come into view.

3: A theory of just educational measurement

> Deciding who advances through what schools and careers is an immense power—one that would have produced furor long ago if a government agency possessed it. The presumption to define what is aptitude and manage the allocation of opportunities based on that definition should be directly challenged.
>
> -Ralph Nader (1979, p. xvi)

> There is a natural aristocracy among men. The grounds of this are virtue and talents... May we not say that that form of government is the best which provides the most effectually for a pure selection of these natural *aristoi* into the offices of government?
>
> -Thomas Jefferson (1813, see: Cappon, 1959, p. 378)

This chapter brings together the lessons Rawls offered about institutionalized measurement with the lessons he offered about the function of educational systems in a just society. Taken together these lessons provide a system of principled distinctions that can be used as a framework for making sense of complex and seemingly contradictory considered judgments about the role of testing in education. As an applied theory of justice it must be able to explain why some uses of testing serve justice while others do not, as well as help structure the redesign of future testing infrastructures. This chapter explicates such *a theory of just educational measurement*. It begins with an exploration of certain basic conceptual distinctions that are needed to get the theory off the ground, including the difference between physical and

psychological measurement and the reductive representation of educational values enabled by testing (referred to as the *education commodity proposition*). Then the discussion focuses on clarifying the complex dynamics of testing-intensive educational reforms, which are classifiable into two ideal types: *efficiency-oriented testing* and *justice-oriented testing*. This sets the stage for Chapters 4 and 5, which offer a series of historical case studies that serve to determine the value and explanatory power of the framework. In order to contextualize the substance of this chapter's discussion, I begin by foreshadowing the historical discussion to follow as well as summarizing what has already been discussed.

Standardized testing: a new moment in the history of institutionalized measurement

> Below are given the names of four animals. Draw a line around the name of each animal that is useful on a farm:
>
> **cow tiger rat wolf**

(*Kansas Silent Reading Test*, 1915. Reported as the first published mass-administered multiple-choice test in: Samelson, 1990)

The question above seems simple enough, although depending where you live there could be tiger, rat, and wolf farms, or a farm where cows play no part. Of course, drawing a line around the word **cow** is the right thing to do. Given its simplicity, it is hard to believe that the idea behind this arrangement of words would change the face of education forever. About a dozen similarly formatted questions were given to thousands of students throughout Kansas in 1915, as a way of determining the literacy rate in the local schools. This was not the first time a test was given broadly among a student population to determine "what was going on in the schools." Horace Mann was known to have mass-administered math tests to all the students in Boston for this reason (Cremin, 1980). But Mann's tests didn't look like this, nor were they scored by the thousands—using stencils that could be overlaid across the answer sheets allowing untrained assistants to eyeball a page and score it in seconds. This was a scientific invention that would soon redirect the lives of countless millions, from army recruits to immigrants, school children, and inmates. Rhetorically backed by scientific theories from the beginning, testing technologies were placed alongside thermometers, x-rays, and precision scales as the fruits of scientific advance made good for public use. Testing would quickly become one of the most lucrative, legally powerful, and expert-rich industries, making a living off the sheer size and complexity of the American educational system (Haney, Madaus, Lyons, 1993; Toch, 2006). Mental testing; psychological measurement; standardized testing; educational assessment; educational measurement—these terms signify a set of relatively new scientific and technological practices, which have had (and continue to have) profound and often unintended social consequences.

Since the end of the nineteenth century, standardized testing practices have constituted a remarkable and still unfolding chapter in the history of institutionalized measurement, and a prime instance of the perennial relationship between measurement infrastructures and social

justice. Following the birth of psychology as a science, there were a wide variety of measurement-intensive research innovations that could have been considered as the foundation of a "science of psychological measurement" (Michell, 1999). But it was one tradition of mental measurement or "mental testing" (a term coined in 1889) that carried the full weight of the new science's ambitions and captured the imagination of reformers, bureaucrats, and everyday people. Fueled by a series of political and administrative necessities, as well as a cultural receptiveness to the solutions of scientific experts, "the testing movement swept the nation as an educational crusade"—a crusade in the name of both social justice and efficiency (Tyack, 1974, p. 207). James B. Conant, a co-founder of ETS (who figures prominently in Chapter 5), described modern testing infrastructures as the greatest "educational inventions" in history, and believed that they ought to be spread with great consequence throughout American institutions just like the other remarkable technologies that had by that time resulted from modern science and engineering. Testing spread quickly on the back of successive waves of technological innovations.

IQ testing was imported to the United States from France and then quickly divorced from its original intended uses. The mass-administrable multiple-choice IQ test was invented in May 1917, transforming a measurement practice once conducted one-on-one for clinical purposes into an industrial-scale technology suitable for "measuring" millions of minds (Samelson, 1990). In 1931 a young science teacher in Michigan invented a means of automated multiple-choice test scoring based on the differential responsiveness of pencil lead to electricity. IBM bought the idea and the resulting efficiencies of scale resulted in the rapid proliferation of fill-in-the-bubble testing infrastructures, beginning a new era of national testing (Lemann, 1999). In the 1980s computer technologies began to transform testing infrastructures once again, this time in terms of

data-analytic capabilities and test-administration procedures. New question formats and large-scale data modeling served to greatly expand the reach, impact, and diversity of standardized educational measurement practices (Haney, Madus, & Lyons, 1993). While many of the newer forms of testing transcended their ancestral technology of multiple-choice, the vast majority of standardized tests were (and still are) largely practices involving the automated scoring of "selected response" items.

Now at the beginning of the twenty-first century, a new internet-based, computer-administered national testing infrastructure is being built. It will include a greater diversity of item formats, but it still leans heavily on institutional practices for measuring minds that were first devised during World War I. This new infrastructure will be composed of high-stakes tests involving selected response items focused on a narrow range of capabilities. Results will be presented using simple hierarchal ranking scales that reduce the quality of performances to as few quantitative indices as possible and outcomes will affect financial and educational opportunities. Educational measurement practices have been remarkably true to their technological and practical roots during the near-century of their existence. They have expanded their scope and impact, been legally institutionalized and ideologically justified in a variety of ways, but the broad approach to the practice of testing has remained remarkably consistent.

Accordingly, few other representations of American education are as iconic as a room full of students sitting in carefully placed rows of desks taking a multiple-choice test. This is a scene that could have easily been witnessed in almost any American school as early as the 1920s, and it is a scene that is equally common today. For nearly a century, standardized testing has played an increasingly important role in an ever expanding system of schools. "The multiple choice test—efficient, quantitative, objective, capable of easily generating data for complicated

statistical analysis—has become the symbol or synonym of American education" (Samelson, 1990, p. 122). These tests have played a wide range of functions, from educational research to the reorganization of whole school systems. They have been used to sort students into capability groupings, inform career counseling, determine college admissions, monitor teacher quality, improve school efficiency, and even quantify the educational achievements of the nation as a whole. Testing shapes the lives of everyone involved in schooling, and the impacts of testing have become increasingly significant as its functions have continually expanded.

Between 1910 and 2010, American public schools evolved into one of the largest and most complex bureaucracies in the world, and like other large bureaucracies it came to require the institutionalization of measurement infrastructures for a variety of purposes. As argued in Chapter 1, measurement infrastructures are a prerequisite for the administration of justice, while at the same time they are deeply susceptible to use as instruments of *injustice*. Standardized testing has been (and continues to be) at the heart of the educational injustices perpetrated by US public schools. Yet testing is (and always has been) closely allied with ideals such as equality, fairness, and justice, which have perennially shaped the discourse surrounding school reform.

This seeming contradiction between testing that serves justice and testing that perpetuates injustice is the central impetus for the philosophical work undertaken here. When an important social practice that effects nearly everyone in society is the focus of intense controversy—with social justice advocates on both sides of the debate—it is time for the kind of philosophical work that Rawls's methods enable. That is, it is time to undertake the principled construction of a philosophical model or theory that serves to clarify our considered judgments and to "define a shared conception of an ideal basic structure... toward which the course of reform should evolve" (1971, p. 231). This is the overall goal of this study, to clarify the relationship between

educational measurement and social justice so that standardized testing infrastructures can begin to be redesigned in light of a principled ethical framework, *a theory of just educational measurement*.

The central elements of a theory of just educational measurement

All institutionalized measurement infrastructures are subject to evaluation in terms of their role in the creation of a just basic structure, leading to the three basic rights of all individuals who are subject to institutionalized measurement practices. It should be clarified here that the principles of just institutionalized measurement apply only to institutionalized measures that function as a part of basic social structures and that directly impact individuals. To be directly impacted by a measurement infrastructure is to be personally involved in measurement events (e.g., one's body, mind, or property is immediately implicated). There are vast arrays of measures that never leave scientific laboratories or that remain tied to scientific activities. These kinds of measures do not function as basic structures and do not directly impact individuals so they do not raise social justice issues. Of course, individuals participate in experiments and run scientific laboratories, but misuse, fraud, and abuse of measures in these contexts fall under the jurisdiction of medial or scientific ethics, not the social justice of measurement *per se*. When scientific measures do leave the lab and impact public life individuals often have an interest in their objectivity and relevance, and may stand to benefit in some way from their proper use (e.g., public-opinion polls or measures of global temperature fluctuation); but the issues raised concern the role of science in society, not the role of measurement *per se*. Some measures, like those used in public health or econometrics, are arguably institutionalized as basic structures because of their standardization, ubiquity, and broad social importance (revealing action-orienting

information about, for example, rates of contagion of a new strain of flu, or indices of impending economic crises). However, these measures do not directly impact individuals, even if they do play a critical role in their decision-making; the issue here is about the ethical and scientific responsibilities of economists and public health professionals, not about the justice of their measurement practices. That said, the lines are not hard and fast, and as soon as a measurement practice begins to directly impact individuals (as in, for example, when a one-time public health measure is used diagnostically), then the three principles of just institutionalized measurement become relevant.

First and most fundamental is every individual's *right to objective measurement*. As demonstrated in the discussion of the unjust miller in Chapter 1, who instituted and manipulated measures by fiat to serve his own ends, a lack of objectivity in measurement practices creates a demonstrably unfair basic structure in which a practice that ought to remain consistent changes depending on who is involved. This almost always results in a situation in which those who have the power to set the measure benefit at the expense of those who do not. Recall the definition of objectivity offered in Chapter 1, where it was suggested that objectivity does not entail quantification or the scientific refinement of measures. A measure is objective when it is demonstrably changeless regardless of context, content, and user bias. So objectivity does not entail quantification, nor does it entail the scientific refinement of measures.

This simple definition allows for certain kinds of qualitative measures to be considered objective, an essential factor when it to comes to educational measurement. It is also important to note that the right to objective measurement is only relevant when the objective measurement is possible and preferable to existing alternatives. Just because something can be measured objectively does not mean it must be, but when individuals are already involved in measurement

practices or see areas of social life that could be better regulated by instituting a measure, they have a right to demand that the measure be objective.

Beyond the foremost right to objectivity, individuals have *a right to relevant measurement opportunities* and *a right to benefit from measurement*. As demonstrated in the discussion of the unjust bureaucracy in Chapter 1, objective measurement infrastructures can be institutionalized in ways that are unjust because they are insensitive to the conditions, values, and needs of those most affected by them. The accouterments of objectivity (scientific authority; technological sophistication; organizational efficiency) can be used to justify disempowering individuals and groups, who when stripped of their metrological autonomy are forced to carry out practices and understand the world in terms dictated by measures that do not meet with their reasoned consent. This often results in irrelevance, alienation, and harm. Therefore all individuals who participate in measurement practices have a right to also participate in the determination of which measures and measurement practices are relevant to their needs. What this means will vary from context to context, but in no context does this imply that each individual gets to pick a measure that serves their private interests or that anyone is free to override the judgment of experts (such as doctors and engineers) who have a special knowledge of certain measurement practices. Rather it implies only that measurement infrastructures are a rightful subject for deliberative democratic decision-making and that institutional innovations are needed (likely to be different in different sectors and contexts) to facilitate the inclusion of all relevant perspectives in decisions about measurement practices.

As discussed below, when it comes to standardized testing these metrological rights are obscured by tendencies to confuse the distinctions between psychological measures and physical measures and to embrace the resulting reductive representations of educational value and related

distortion of social relations (e.g., the terms of the education commodity proposition). The inescapable normativity of testing instruments requires that they be treated as more than just another institutionalized measurement infrastructure. The general metrological rights outlined in Chapter 1 must be integrated with a broader set of philosophical commitments about the nature of educational institutions, otherwise the distortions of educational relationships produced through testing cannot be subject to analysis.

These additional principles were provided in Chapter 2 where it was argued that an educational system in a just society must fulfill at least three functions. The first function is to assure everyone access to the educational primary goods they are entitled to as members of society. Educational primary goods include those experiences and opportunities that reliably generate both the competencies necessary to participate in political institutions as well as those needed to pursue a self-chosen conception of the good life. A school that can provide this kind of education requires a certain social organization and curricular content, both of which must embody the broader commitments of society to social justice. The second function of educational systems in a just society is to contribute to a basic structure that provides for fair equality of opportunity. This positions schools as part of a network of economic and political institutions which must work together to assure that careers are open to talents and that opportunities for positions are allocated through procedures that are fair to everyone involved. And finally, schools must make possible the unique self-actualization of each individual as specified by the demands of the *Aristotelian principle*, which posits that all humans have an inborn desire to master complex skills and to exercise them in contexts over which they have control.

These are the central elements of a theory of just educational measurement—three metrological rights and three educational commitments:

Principles of just institutionalized measurement

The right to objective measurement: individuals have a right to objective measurement wherever possible and wherever preferable to existing methods.

The right to relevant measurement: individuals have a right to participate in the determination of which measures are relevant to their needs.

The right to benefit from measurement: individuals have a right to benefit from measurement practices that directly involve them.

Principles of just educational institutions

Educational primary goods: educational systems should provide for a certain "amount" of education that is owed, as a right, to all members of society; the scope of these educational primary goods is to be determined by the principles of justice governing society.

Education for fair equality of opportunity: educational systems should be designed as an essential part of a system of institutions (a basic structure) that is built to ensure fair equality of opportunity for everyone.

Education for self-actualization: educational systems should promote the development of self-actualized individuals (who have complex cognition, mature socio-emotional lives, and satisfying work).

Taken together these two sets of three yield a matrix of conceptual distinctions that can be utilized to characterize the ethical complexities involved in testing practices. The analytical power of the theory resides in the inter-animation of these elements as they play out in the dynamics of testing-intensive educational reforms. The most common such dynamic is discussed below as *efficiency-oriented testing*, which includes an even more basic set of explanatory terms, known as *the education commodity proposition*. But underlying all these dynamics are a set of assumptions about the nature of the knowledge produced through testing. These are the epistemological assumptions underlying testing-intensive educational reform, and they must be understood before any theory of just educational measurement can be put to use.

Between the physical and the psychological

> The term "normative facts" has been happily introduced into the general vocabulary [as distinguished from a "natural or physical fact"]…to describe that which constitutes a norm for the subject and, at the same time, an object of analysis for the observer engaged in studying both the behavior of the subject and the norms he recognizes…. [For example] normative facts are studied in developmental psychology when the question is to discover how subjects who were originally insensitive to certain logical norms come to regard them as essential [i.e., "learning as the transformation of *how* one thinks, not as mere additions to what one knows"]...
>
> -Jean Piaget (1970, p. 8-9)

Understanding how some obvious injustices perpetuated by testing escape the view of well-intentioned reformers requires examining certain epistemological assumptions about the nature of measurement technologies. These epistemological assumptions underlie the education commodity proposition and thus facilitate the hegemony of efficiency-oriented testing practices. As discussed in the next section, the educational commodity proposition depends in part on the power of metaphors that blur the distinction between testing and physical measurement. The problematic social relationships it engenders thrive parasitically on a basic set of epistemological confusions about the differences between physical and psychological measurement. Clarifying these important differences is a necessary precondition of a coherent theory of just educational measurement.

Physical measurement infrastructures involve people, but they are not primarily about people. Even physical measures applied to people (e.g., measures of height, weight, and blood pressure taken by doctors) are applied to the body as if it were a clinical *object*. In fact, as explored in the next chapter, one of the most common ways of placing psychological measurement practices under the rubric of physical ones is to use the language of *medical diagnosis*, characterizing standardized testing as akin to either clinical diagnostic practices (like the placing of a thermometer under the tongue) or public health and disease prevention (like health screening and vaccination programs). The other metaphor that has been widely used to frame large-scale testing is *social engineering*, which makes testing akin to the complex measurement practices required to build a bridge or engine. While this term has gone by the wayside, the centrality of the analogy has not—standardized testing is still often conceived as a tool that can be used to determine the "functional fit" of individuals within institutions, to assess "human capital," and "channel manpower" into the broader social system.

The history of testing (explored in coming chapters) is in part *a history of failed metaphors*. It is in part a history of well-intentioned but ill-conceived endeavors to do with psychological measures what had been done with physical measures in medicine and engineering. These ways of conceiving standardized testing infrastructures simply do not fit with the realities of how testing practices are actually undertaken. As will be shown, these metaphors provide a language and a set of models that easily accommodate the terms of the education commodity proposition and efficiency-oriented testing.

Psychological measurement practices produce a unique kind of knowledge. Standardized tests are based not on the reliable differential responsiveness of physical instruments to physical realities, but on the objective and standardized treatment of "normative facts." This term was coined by the great Swiss psychologist Jean Piaget (see: Smith, 2006), a student of Theodore Simon, who was a close colleague of Alfred Binet, the inventor of the IQ test who is discussed at some length in Chapter 4. The difference between normative facts and physical facts can be easily seen in comparing a ruler or thermometer to a standardized test. A ruler or thermometer is made up of units that are value-neutral and come to have a normative component only when they are used in certain contexts. For example, a fever of 104 is bad, but that is a great temperature for my soup; a board that is two feet long is useless for this project, but perfect for a different one. In and of itself, determining that an object is a certain temperature or length has no normative content; it is a simple physical fact (if ever there were such a thing). It is precisely this power of measurement techniques to create such simple incontestable facts that puts advances in measurement at the center of nearly all scientific advancements, especially scientific revolutions (Kuhn, 1962). It is because measures are ostensibly "value-free"—their units do not come

predetermined with normative connotations—that they are able to be used as a kind of theoretically neutral bedrock upon which various theories can be based and tested.

Standardized tests, on the other hand, are *always already* value-laden because they are constituted by a series of items with "right" and "wrong" answers (or by a series of rubrics that mark a gradation from "worse" to "better" answers). In and of itself, receiving a certain score on a test has normative content. This evaluative dimension is deepened and augmented by context, but it is not created by context as is the case with physical measures. Testing infrastructures deal in *normative facts*—not facts about the properties of physical things, but facts about properties of human judgment, such as what is "correct" and "incorrect." Therefore a test is by definition an evaluative judgment about a person, even when it is completely objective, standardized, and quantitative. The units and instrumentation are value-laden by there very nature. Believing otherwise requires a conceptual sleight-of-hand (typically based on one of the disanalogies mentioned above) where the test is understood as simply revealing "what is the case," as if its function was merely descriptive and not simultaneously also evaluative. Overlooking this (and imagining that a test is as unproblematic as a physical measure) can obscure understandings of the impact tests have on the individuals who take them.

Even the analogy to medical diagnostic measures (which appears to account for the normative dimension of testing) is based on this epistemological mischaracterization of testing. Medical diagnostics seem to be intrinsically value-laden only because they are used repeatedly—and often solely or primarily—in a very specific context. It seems as if a blood pressure reading of 140/90 mmhg (as indicated by a precisely calibrated and universally standardized blood pressure gauge) is always a bad thing. But in fact using that same gauge and getting those same numbers would not be a cause for alarm when taking the blood pressure of an animal such as a

dog. Make no mistake, the claim here is not that biological realities do not have normative properties—high-blood pressure is bad, so are the heart attacks and strokes that often accompany them. When a human gets a blood pressure reading of 140/90 mmhg it is (almost) always a bad thing, given the nature of the human organism. The argument here is rather that the measures used to characterize the state of biological organisms are not constituted by or made up of normative distinctions. Despite common associations, medical instruments used for diagnostic purposes are not intrinsically normative in the way that standardized tests are. They are not normatively structured; the units and instrumentation are not value-laden—they receive their evaluative meaning entirely from the context in which they are used. And while tests do receive additional evaluative meaning depending on the contexts in which they are used, the very act of measurement itself is always already value-laden—the test is intrinsically normatively structured.

This distinction between normatively structured (value-laden) measures and non-normatively structured (value-neutral) measures is an important part of many post-positivistic approaches to the philosophy of science (Piaget, 1970; Habermas, 1988; Bhaskar, 2013). Following the failure of philosophers to build a single unified theory of science (bringing physics, biology, psychology, and sociology under the same broad meta-methodology), many philosophers began to argue that the methods and practices of the human sciences are *irreducible* to those of the physical sciences. Most adopting this stance have been quick to argue that this differentiation does not entail that the human sciences are any less scientific or objective—quite the opposite. As Habermas (1988) has argued, it is by respecting the differences between the human and the physical sciences (not through their conflation) that the human sciences can attain their proper epistemic standing. And the heart of the difference resides in the fact that the

methods and measures in the human sciences require the establishment of an *interpersonal relationship* with the subject under investigation, whereas the physical sciences do not.

Normatively structured methods and measures—that target human judgments and decisions—transcend but include investigative methods based on mere observation. It is, of course, possible to observe the results of a judgment or decision (e.g., I can easily observe you going to a school board meeting), but mere observations of behavior do not enable an understanding of judgments or decisions that cause it (e.g., why did you decide to go?). Understanding a judgment requires taking up the role of a *participant observer*, sharing first in a mutual understanding of the rules and beliefs constituting the judgment before then making this interpersonal reality into an object that can be classified, explained, or otherwise re-described scientifically.[4] This means that understanding a judgment requires making a judgment. For instance, I have not understood what you mean when you say "schools are the most important institutions in a democracy" if I do not know what it would mean to agree or disagree with you (Habermas, 1988.). This is even clearer in the case of a multiple-choice test item, where human judgment is distilled into a forced choice that is understood entirely in terms of its being correct or incorrect. A choice of "A" is meaningless without making certain assumptions about the

[4] This view implicates issues related to the *explanation* versus *understanding* debate, which has been an important part of the philosophy of the social sciences from Dilthey and Weber through Popper, Hemple, and Habermas. The debate revolves around two distinct ways of treating human action scientifically. The paradigmatic case is of actions that can be *explained* in ways that undermine or contradict how the actors themselves *understand* them: e.g., churchgoers understand their actions in terms of a belief system, and social scientists can work to understand churchgoing actions according to the beliefs explicitly held by churchgoers; but social scientists can (and often do) bracket the explicit beliefs of actors and explain churchgoing in terms of certain social functions, like group cohesion or ideology, which could be (and often are) denied as a motive for action by the actors themselves. Neither approach is right or wrong, but I hold the view that it is highly desirable that the social sciences pursue adopting *both* approaches, and that understanding is often a condition for the possibility of good explanations. The basic argument behind this more fundamental claim is that attempts to explain an action without first understanding it run the risk of mischaracterizing the very thing being explained: if a man goes into a store we must know if he understands himself as needing to shop or as needing to kill time, two actions that look the same, but are, in fact, different, an thus would have very different explanations. There is not space to fully justify this view here, which is held by Apel (1984) and Habermas (1988), among others (Von Wright, 1971).

judgment that led to this choice, e.g., that it was not a guess or an involuntary movement that led to the selection. Tests cannot be built, administered, scored, or interpreted without a tacit reliance on shared cultural norms, rules of inference, and a range of accepted truths about the linguistic and cultural world. There is no way to get "behind" or "around" the normative dimensions of psychological measurement; unlike a ruler, thermometer, or scale, a standardized test is by its very nature a set of evaluative distinctions.

This means that every testing event requires, through its very structure, the establishing of an interpersonal relationship. No matter how or where it is administered—from across a table or over the internet—a test institutes a relationship with unique properties, based most fundamentally upon a communicative dynamic of performance and evaluation. Every testing event creates a relationship, and every relationship created through testing is one in which there is: 1) a differential of power between administrator and test-taker; 2) a preexisting system of evaluative categories in terms of which the test-taker is re-described as a social object; and 3) a cultural setting in which the *effects* of that power differential and that evaluative system reverberate. None of these are apparent if testing is understood as a non-participatory physical act or as a mere observation, which ignore the epistemic status of testing as a social activity. A full discussion of these epistemological issues is beyond the scope of this work (but see: Stein, 2007; Stein, Connell, & Gardner, 2012).

The point here is to mark an important distinction between physical measurement and psychological measurement in order to clarify the uniqueness of the social justice issues involved in testing practices. The intrinsic normativity of psychological measurement instruments is an important part of what gives structure to the interpersonal relationships that are engendered through testing. The education commodity proposition and efficiency-oriented testing (discussed

below) are both the necessary and rational outcomes of advances in objective testing, yet they are also prone to excess and can come to perpetuate injustice through overuse and overemphasis.

All institutionalized measurement practices raise a common set of social justice issues—irrespective of whether or not they are physical or psychological. These are addressed by the principles of just institutionalized measurement outlined in Chapter 1. This section here has been intended to make it clear that psychological measures implicate an even broader set of social justice issues than physical measures do. These additional concerns deal with the inevitable impact of testing infrastructures on the interpersonal relationships and cultures in which they are embedded and the goals of the educational institutions in which they are deployed. These broader concerns mark out unique territory for testing in the metrological landscape.

The education commodity proposition: tests as the coin of the educational realm

> Tests in the US are now used to make a range of important decisions about individuals and institutions.... [In many] instances the decisions have the force of law…. Tests are being used for certification or recertification of teachers, promotion of students from one grade to the next, award of high school diplomas, assignment of students to remedial classes, allocation of funds to schools and school districts, award of merit pay to teachers on the basis of their student's test performance... and the placement of school systems into "educational receivership." More generally, tests are now widely being touted as instruments of

> national education reform and renewal. In short, tests are becoming the coin of the educational realm.
>
> -Haney, Madaus, & Lyons (1993, p. 57-58)

Standardized tests are unique in the history of institutionalized measurement for a variety of reasons, most of which implicate social justice as a predominant concern in their design and implementation. As discussed in Chapter 1, the predecessors of physical measures were devised early in humanity's prehistory. While there was great diversity in practice involving a wide range of approaches to instrumentation and standardization, there were a relatively limited range of properties that were subject to large-scale standardized measurement: length (from small dimensions to long distances), time (from subdivisions of days to calendrical record keeping), weight (from weighing gold dust to weighing livestock), and volume (from sacks and bushels to cart loads).

As discussed in Chapter 1, most measurement practices began as communal solutions to recurring situations that required the coordination of social action with specific aspects of the world. Most measurement practices began as inventions spawned by social necessities and evolved to become part of the taken-for-granted background of practices and institutions that constitute a society. Measurement infrastructures become basic structures. As will be explored in Chapters 4 and 5, the history of testing can be recounted through this kind of narrative in important ways. Psychological measurement infrastructures emerge and flourish for many of the same reasons that physical measurement infrastructures do. They also tend to fade into the background to become part of accepted and taken-for-granted social structures. However, this analogy—that psychological measurement is akin to physical measurement—can be

overextended, and doing so has resulted in fundamental misunderstandings about how standardized tests can and ought to be used.

The discussion in the previous section demonstrated that the social justice issues involved with educational measurement transcend but include those involved with physical measurement. Simply put, the main way in which psychological measurement differs from physical measurement is that *people's minds* are the subject of psychological measurement, whereas specific properties of objects and processes are the subject of physical measurement. Two sets of issues related to this seemingly simple fact require attention, one set is epistemological (already dealt with in the previous section), the other is social and ethical. The goal here is to begin to consider the role of testing in the dynamics of social systems, determining how it can come to be just or unjust. The full-blown dynamics of testing-intensive educational reforms are discussed in the next section. The central issue here—referred to as *the education commodity proposition*—requires some elaboration first.

The education commodity proposition appears in institutional cultures as a simplification of decision-procedures comparable to similar simplifications that accompany what political economists call the "labor commodity proposition" (Bowles & Gintis, 1986). One of the foundations of the capitalist economy is the idea that individuals are free to sell their labor as a commodity that is fundamentally no different from other commodities, being part of a marketplace with price competitions governed by supply and demand. The labor commodity proposition—that labor is just like any other commodity—justifies, among other things, the persistence of efficiency-oriented decision-procedures and data-driven management strategies in which labor is represented as if it were simply another purchasable component in the production process: to be sought for as cheaply as possible and utilized with maximal efficiency.

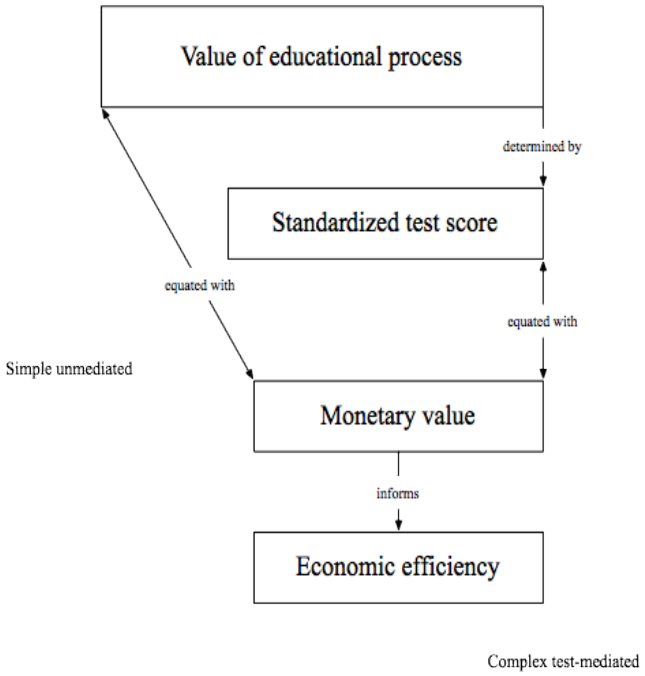
Of course, labor is actually unlike other commodities (such as a TV or car), because labor is *inalienable* from the person who "sells" it. When I sell you my TV in exchange for a sum of money, you walk away with the TV and I walk away with the money. But when I sell you my labor, I must live through the work to be done. The valuation of my labor (assigning it a numerical, cash value) and my fulfillment of the employment contract (exchanging my labor for what it is deemed to be worth) implicates me both physically and psychologically in a way that the exchange of my TV for cash does not. The analogy between labor and other commodities breaks down once the economy is understood not merely as a system of exchange, but also as a system of employment—it is not just *things* that are caught up in market dynamics, it is also *people* (*ibid*.). Framing economic issues in terms of the (dis)analogy "labor is a commodity" removes considerations about the human side of the labor-exchange relationship, simplifying its representation in terms of monetary units, and depoliticizing it by turning it into an impersonal relationship—like the selling of a TV—to be negotiated and determined in terms of what the market will bear (as opposed to what justice requires due to the involvement of *persons*).

An oversimplification similar to the labor commodity proposition accompanies the use of large-scale standardized testing infrastructures. It is this oversimplification that is referred to as *the education commodity proposition*. Testing can be used to put a number on a "unit" of education, giving it a quantitative value and making it amenable to certain kinds of decision-making calculations, such as cost benefit analysis and the estimation of returns on investment. That is, testing allows educational processes to be conceived as if they were no different from other commodities, simplifying their representation to monetary transactions that can be considered in terms of economic efficiencies. This does not necessarily contradict the role of testing in the science of education (a point returned to below and in the conclusion), but it does

mean there is a parallel use for a great deal of testing-intensive educational research, e.g., tracking, saving, and making money (Ravitch, 2013).

Figure 1 shows the ideas that constitute the education commodity proposition. In its simple unmediated form the education commodity proposition simplifies the value of an educational process down to a price. The idea that the value an education can be represented by is its price was one of Socrates' primary concerns in his debates with the Sophists, who he claimed were distorting the value of education by adapting it to be sold to the highest bidder. But it is the complex test-mediated form of the education commodity proposition that is the focus here. In this case the value of an educational process is reduced to test scores, which are then equated with monetary values and factored into calculations of economic efficiency.

Figure 1: *the education commodity proposition*



In its fullest sense the education commodity proposition should not be understood as necessarily requiring test-based quantification (although test-based forms of the education commodity proposition will be the focus of the discussions that follow), nor is test-based quantification always undertaken for the purposes of representing educational value in financial calculations. There are many ways to translate the value of education into monetary terms that don't involve testing or even quantification, as in for example, the use of exit interviews, testimonials, and case studies, which seek to convey the "return on investment" of education in qualitative terms. Likewise, simple non-test-based forms of quantification, such as graduation rates and employment statistics, can be and are utilized in calculations of the cost effectiveness of educational programs. On the other hand, the test-based quantification of educational value is not necessarily always motivated by financial concerns. There are many reasons to quantify educational processes aside from needing the numbers for use in economic calculations. For example, in a situation where money is no object educators will still have reasons to employ objective indices of program effectiveness; test-based quantification plays a valuable role in program evaluation studies, even when the cost of the program is of no concern. Moreover, in situations where money matters, not every use of test-based quantification will be implicated in financial calculations.

So there are three distinct steps that get us to the "logic" of the education commodity proposition that is the focus here—e.g., the complex test-based form. First is a desire to place a financial value on educational processes, which can be done with or without test-based quantification. Second is a desire to quantify educational value through the use of standardized tests, which can be done for economic or non-economic reasons. Third is a decision to meet the

desire for placing a financial value on education (step 1) by employing test-based quantification strategies (step 2). This third step, which combining specific variations of the prior two, leads to the "logic" of the education commodity proposition discussed below.

Consider one of the perennial questions testing infrastructures are used to answer: *how much* education are we getting for our tax dollars? Or, more recently: *how much* does this teacher measure up in a value-added analysis? These questions are based on a strong (implicit) analogy between the nature and functions of psychological measurement and those of physical measurement, as if the object of measurement was *the amount of some thing*. Conceiving psychological measurement this way obscures the *relationships* engendered through testing (e.g., between student and teacher, teacher and administrator, administrator and government) in similar ways as the labor commodity proposition obscures the relationships involved in the employment contract. This results in the apparent de-politicization of the relationships produced through testing because they are seen as being based on the kind of unproblematic relationships involved in physical measurements, e.g., impersonal relationships between a person and an object by way of a measurement instrument. It also results in the reduction of questions about justice to questions about efficiency, as discussed in the next section. With a problematically test-dominated educational culture in place, institutional relationships are further misconstrued and then misunderstood by those involved as being less complex and ethically fraught than they actually are.

As with labor in the labor commodity proposition, education is not simply reducible to the terms of market exchange because it is *inalienable* from the individual being educated. Individuals are not given an education in the same way they are given a TV or some cash. Individuals *become* educated; they are shaped by the total experience of whatever educational

processes they participate in. This is obscured by the education commodity proposition, where testing is used to simplify how the value of the educational process is represented and thought about (and often how it is literally *calculated*). Just as the value of labor can be reduced to its cost, allowing the relations of employment to be governed only by what the market will bear, so the value of education can be reduced to standardized test scores that are then converted into monetary terms, allowing the relations of teaching and learning to be governed solely by the demands of economic efficiencies. The "logic" of the educational commodity proposition dominates decision-making in many spheres.

On the "logic" of the education commodity proposition

The discussion in this section is not intended as a presentation of facts about the existing educational system, or as a sociological and economic analysis, but rather as a demonstration of the "logic" resulting from the terms of the education commodity proposition. While much of what is discussed does go on in schools today, the goal here is only to "run the model"—to explore the implications of thinking about education according to this powerful and simple model in which test scores are converted into financial terms. The presentation is thus one-sided, presenting the most basic thought-forms spawned by the education commodity proposition. This does not imply that these ways of thinking and decision-making dominate all educational organizations, or that there are not important forms of resistance. Critical reflections and observations referencing the current state of education are offered along the way, but the "facts" about how the education commodity proposition has impacted schools are presented in Chapters 4 and 5. The main goal here is to display the reasoning behind the education commodity

proposition. This way of thinking has created a powerful and seemingly irreplaceable function for testing in many educational configurations. The education commodity proposition influences a great deal of educational policy because it is a straightforward and seemingly necessary way to think about schools in economic terms. It appears necessary mainly because economic necessities bear down strongly on so many school systems, and many educators have no choice but to foreground their financials. This trend is intensified by federal policies that have school compete for funds based on test scores and by the fact that an increasing number of schools are being run like businesses (with many set up as for-profit ventures). This sense of necessity, which increases the prevalence and power of test-based forms of the education commodity proposition, constitutes part of the "logic" unfolded in what follows.

The clearest examples involve those who *invest* in educational processes: governments, philanthropies, venture capital. Simply put, if the *amount* of education you are getting for your investment is represented (only or predominately) by the numbers generated on tests, then moving these numbers becomes the only way to "see" those changes in the value of the educational process that are the intended result of the investment. If you do not measure it you cannot monetize it, and if you cannot monetize something then you cannot technically "see" if an economic investment in it has worked.[5] This summarizes the main problem facing governments that invest tax dollars in public education and are then required to demonstrate that this public money was well spent—the watchword here is "accountability." In these situations test scores are used to translate (or re-represent) the value of educational processes in economic terms. The

[5] I am working with a simple notion of what it means to monetize something, based the definition found in most dictionaries: to convert something into a form that can be expressed as currency. So my claim is that testing converts educational value in to something capable of being expressed as interchangeable units that are linearly ordered and quantifiable, that is, units that take a form like that of money, which has just such a quantitative structure. This use does not imply the primacy of profit, privatization, or marketization, which rely on this form of monetization, but are not necessitated by it.

same kind of decision-procedures are used by philanthropic donors and venture capitalists, both of whom must demonstrate due diligence when putting money toward the improvement of education—and improvements can only be monetized if they can be measured in ways amenable to quantitative demonstrations of return on investment.

Of course, the education commodity proposition also frames the decision-making of those who are responsible for organizing educational institutions—the school leaders who are concerned about their own budgets and the effectiveness and efficiency of their internal policies. For example, the value of a new math curriculum is easily turned into a question about the relationship between its cost and the test score gains that result. If the math curriculum that produces the best scores is too expensive, than the next best affordable option will be chosen, often irrespective of other salient differences between the two curriculums. Curricular decision-making becomes guided by the calculus of cost-benefit analysis, which removes a wide variety of complex considerations from the table by distilling the problem down to its "essence"—economic efficiency. More importantly, suppose this next best curriculum produces test score gains but causes students to dislike math, creates misunderstandings not detected by the test, or results in teacher burnout—just some examples of differences in educational value that easily and regularly escape measurement and therefore are not included in the official calculation of what the curriculum is "worth."

It is clear that the education commodity proposition shapes the decision-procedures of those who *invest* in educational institutions (e.g., governments and philanthropies) as well as those who serve as administrators in them (e.g., school leaders). But its impact extends beyond those who are in these positions, affecting nearly every individual with a stake in the educational system. On the one hand, it shapes the thinking of those who provide educational services—

those who "*sell*" education. This group includes teachers, who exchange their educational labor for a paycheck, as well as organizations such as colleges, independent schools, and tutoring/test-prep companies that provide educational goods in exchange for a fee. On the other hand, the educational commodity proposition also shapes the thinking of educational *consumers*, namely students, who exchange their time and money for educational goods, sometimes voluntarily (e.g., higher education) and sometimes not (e.g., children are mandated consumers of education, which they often "get for free" in public schools—stretching the analogy of student-as-consumer and the education commodity proposition too far, as will be discussed below). Importantly, students are also the *product* of educational processes, and thus embody the educational value provided to them by the institution they attend. This complex dual role of being both a consumer and a product puts students in an extremely vulnerable position, as testing-intensive determinations of educational value come to dominate not only their choices and actions, but also how they are treated by teachers and administrators, which fundamentally impacts their self-understandings.

Teachers occupy a unique position in school systems because they can be understood (according to the terms of the education commodity proposition) as selling a distinct type of *educational labor*. That is, they exchange their ability to teach for the means of their livelihood. This is important because above all other educational goods potentially conceived as commodities (from books to computers) it is *teaching* that has the most potential to add value to educational processes. It is also one of the most *expensive* aspects of schooling. Therefore teaching is (and has been since the late nineteenth century) one of the central focuses of testing-intensive determinations of the value of educational processes. Yet, teaching is, like all forms of labor, *inalienable* from the teacher. Thus unlike improved school buildings, lunch programs, or technologies, determinations of the educational value added by a teacher concern the work done

by an individual, who can adapt the nature of their work to the methods used to determine its value. Teachers often come to understand their own work according to the measurement categories used to determine the value they bring to the educational processes in their classrooms.

The problem of determining the educational value added by individual teachers easily lends itself to simplification through cost-benefit analysis via testing: two cheap inexperienced teachers whose students show moderate gains can be hired for the price of one veteran teacher whose students show gains that are only slightly larger—thus controlling costs while providing for twice the number of students. Of course, there are countless *undetected* forms of educational value provided by teachers that are not factored into such calculations, such as their ability to foster independent thinking, deal with interpersonal struggles, or bring hope and humor into the lives of their students. Likewise, bad teachers can raise test scores while offloading the collateral damage done by their pedagogy into the *unmeasured* facets of students' lives, by creating toxic levels of stress, promoting mind-numbing test-prep approaches to learning, or by simply focusing attention only on those students whose improvements are key to raising the aggregate class score (e.g., ignoring the top of the class, who will score well regardless, as well as the bottom, who will score poorly even if they show gains).

This is how the education commodity proposition comes to structure the thinking and actions of teachers: *they adapt their teaching to the terms used to quantify it*. The full range of educational values made possible through a teacher's abilities cannot be represented in terms of test score gains. So teachers put their energies toward those values that can be. While tenure and seniority can limit the impact of this approach to quantifying educational value, they do not eliminate it. This is because both students (and their parents) and administrators often perceive a

teacher's value according to the terms of the education commodity proposition, even if the teacher does not (Ravitch, 2013).

Students themselves are subject to the terms of the education commodity proposition in more complex and ethically fraught ways than any other participants in the educational system. Firstly, according to the terms of the education commodity proposition students are cast as consumers of educational goods who are free to make complex judgments about the value of the institutions they choose to enter. Of course, younger students technically have no choice because their parents or guardians choose for them as proxies, but the idea remains that families have an interest in making informed decisions about the value of their educational options. Because school systems are routinely ranked according to test scores, testing success often determines what neighborhood or suburb a family chooses to live in (and the more affluent the family the easier it is to make that educational criterion a top decisive priority). Those families who are not free to choose where to live are free to participate in appeals to improve the goods offered by these schools. However, because any proposed programs require increases or reallocations of funding, debates about their value are likely to be reduced to questions about test-score gains. Suggested improvements that do not lend themselves to test-based quantifications of value are not likely to be carried out unless a program can be somehow characterized in those terms (e.g., music classes will improve scores on math tests). Existing programs that cannot prove their value in those terms (regardless of the other values they bring) are likely to be the first programs cut when funds are tight. The terms of the education commodity proposition create an atmosphere in which the impact of "consumer advocacy" is limited by the metrics available to monitor the quality of the products.

Private school enrollments are not a matter of geographical proximity and these schools are free to pursue a variety of direct-to-consumer advertisements. Educational marketing strategies regularly include test scores or test-score-mediated college-placement statistics. Sports programs and extra curricular activities can also be represented in ways that quantify their value, usually in terms of college placements and scholarships. College placement itself can be represented in terms of differentials in probable future earnings. Likewise, other privately run educational providers, such as tutors, test-prep services, and educational consultants aim to attract potential customers with statistics about test-score gains and resulting downstream financial benefits. Dissatisfied customers are free to go elsewhere, but their choices are still tied to whatever means are available to put a number on the value of the education being provided. The simplicity of these representations, the ease of integrating them into economic calculations, and the lack of trusted alternatives combine to create an environment in which it is difficult for students to understand their options in any other terms.

Of course, regardless of where students go to school and the range of options that are open to them, in the US they are forced by law to spend a large portion of their waking hours in schools until the age of 16 (or 18 in some states). Even at the college level, where students are finally free to choose for themselves where to go or whether to go at all, the economic drawbacks of not attending college are perceived to be so great that it has become a kind of forced choice; those who do not attend typically describe themselves as unable (usually for financial reasons), *not* unwilling (Darling-Hammond, 2010). This creates a marketplace for education that is unlike the marketplaces that exist for other goods, which are not predicated on legally mandated consumption or long-term negative drawbacks resulting from a failure to consume (notable exceptions are the healthcare and insurance industries). The contemporary rhetoric concerning

"school choice," which leans heavily in favor of a privatized school system that functions more like a "true marketplace," tends to ignore the fact that the vast majority of marketplaces are entered and exited voluntarily (Ravitch, 2010). This is a point that will be returned to throughout: educational primary goods are not commodities, they are entitlements—access to these goods is *a basic right*—and their just dissemination requires decision-procedures that transcend but include those emphasized by the education commodity proposition.

As discussed in Chapter 5 with reference to current trends in federal reforms, the idea of school choice and the related analogy of the student-as-consumer obscure the fact that students are the *product* of the educational system, not merely customers in the market for educational goods. Test scores are intended to objectively represent students' abilities as well as changes in students' abilities (i.e., learning as test-score gains). The terms of the education commodity proposition literally represent students' psychological lives as objects with quantitative properties that can be monetized. When an educational process is evaluated in these terms students are understood as an "outcome" or "product" of the educational process. Students are the depositories of the value added by investments in a new curriculum or teaching staff—it is in their skills and capabilities where the value-added measurement must be detected. If more money is put into a school then better students ought to come out. In order to know *how much* better the students are their intellects must be rendered quantifiable. This way of thinking characterizes students as a kind of "raw material" that is worked over as it makes its way through production processes. This is a point returned to in future chapters in the context of discussions about the dominance of the *human capital theory* as a framework for educational reform.

Trust in numbers: the educational commodity proposition and cost-benefit analysis

Make no mistake, the argument here is not that the education commodity proposition and the forms of thought it engenders are wrong and should be done away with—in fact, they are often a correct and necessary part of educational thinking. If the government or a philanthropic organization invests in a new curriculum that will impact the lives of millions of teachers and students, it is clearly necessary to investigate if it is improving their teaching and learning. It is hard to reasonably deny this. It is also hard to deny that the means used to do such an investigation should be as "objective" and "scientific" as possible. The argument is not that this way of thinking is incorrect, but rather that it constitutes a *true but partial* way of understanding the value of educational processes, and it therefore should not be the primary orientation guiding the institutionalization of testing infrastructures. The value of education is not reducible to the terms of the education commodity proposition without remainder. This irreducibility is related to the irreducibility of questions about justice to questions about efficiency, which forms the central argument of the theory of just educational measurement being built in this chapter.

Even assuming that the tests used in making decisions according to the terms of the education commodity proposition are objective, valid, and reliable (which is often not a safe assumption, as will be discussed later)—and assuming that the calculations based on them are thus "correct," as far as they go—it is clear that they distort the representation of educational value by neglecting the unmeasured (and *un-measurable*) aspects of the educational process. No school leader is blind to the multi-dimensionality of the value teachers bring, yet the terms of the educational commodity proposition come to dominate their decision-procedures because they would be radically vulnerable to criticism if they based crucial decisions on other ostensibly "subjective" criteria. When you are making decisions about how to spend money that was given

to you (by tax payers; venture capitalists), you cannot base decisions on your "opinion" (no matter how well-informed or judicious), let alone ask those who gave you the money to "trust you," which is how non-quantitative decision-making is viewed in large bureaucracies.

Theodore Porter (1995, pp. 189-196), an authority on the history of quantitative objectivity, explains this simple truth about decision-making in bureaucracies, especially public ones, where "mere experience or know-how is not sufficient to ground public expertise [because the public] insists that administrative decisions be depoliticized. The ideal is a withdrawal of human agency.... Subjectivity creates responsibility. Impersonal rules can be almost as innocent as nature itself:

> Cost-benefit analysis was intended from the beginning as a strategy for limiting the play of politics in public investment decisions. In 1936, though, army engineers [who invented the decision-procedure] did not envision that this method would have to be grounded in economic principles... or that such regulations would require [the standardization of measurement procedures] throughout the government and be applied to almost every category of public action. The transformation of cost-benefit analysis into a universal standard of rationality... cannot be attributed to the megalomania of experts, but rather to bureaucratic conflict in a context of overwhelming public distrust. Though tools like this can scarcely provide more than a guide to analysis and a language of debate, there has been strong pressure to make them into something more. The ideal of mechanical objectivity has by now been internalized by many practitioners of the method, who would like to see decisions made according to a routine that, once set in

> motion by the appropriate value judgments on the part of those politically responsible and accountable, would—like the universe of the deist—run its course without further interference from the top.

When turning to cite contemporary examples of the expansion of these decision-procedures, Porter (1995, p. 198) looks to educational bureaucracies, quoting from a prescient 1994 front-page article in the *Los Angeles Times* about the plans of American college accreditation agencies to measure how much education college students are receiving. "There is a very significant body of opinion in higher education," stated the accreditation agencies' representative, "that says to the public, 'Trust us. And don't require us to produce any evidence [of results].' What we are saying is that those days are over." Porter reflects on this drive toward accountability as a form of mechanical objectivity enabling the quantification of value for bureaucratic purposes. "Like every institution, the university must be refashioned as a panopticon to open it to surveillance by law courts and regulatory bureaucracies... subsuming its activities within a culture of evidence." And a culture of evidence is almost without exception a culture of measurement. Again, it is hard to reasonably deny the importance of this way of considering education. Why wouldn't we want objective evidence that public money and private tuitions create colleges that get results? It is precisely the simplicity and the rationality of these justifications for the education commodity proposition that makes it so prone to misuse and overuse. It is such a powerful analytical tool that questions are almost never raised as to when it should *not* be used or about how its use should be limited and contextualized.

Importantly, the education commodity proposition begins with insights into the inefficiency and potential injustice that comes with a lack of objectivity—reflecting the first

principle of just institutionalized measurement articulated in Chapter 1. "For practical and moral reasons alike, efficient democratic government seems to require improved methods of accounting, statistics, and other forms of quantification" (Porter, 1995, p. 152). The social justice issues surrounding the educational commodity proposition are thus not those that stem from a *lack* of objectivity. The educational commodity proposition results in injustices that are due to an *excess* of objectivity. Taken to extremes, objectivity and its institutional accruements (e.g., power invested in external experts and technologically enabled quantitative reasoning) will begin systematically to distort the educational process, eventually overriding the autonomy of students, teachers, and administrators in unacceptable ways.

This brings into play the second principle of just institutionalized measurement articulated in Chapter 1, which focuses on *who decides what gets measured and the relevance and benefits* of the selected measures to those who are immediately subject to them. Recall the lessons of the unjust bureaucracy, in which efficiency experts violated the metrological autonomy of individuals and destroyed the integrity of their practices, replacing them with a newer and better one, which then (ironically) resulted in failures of the enterprise. Violating individuals' metrological rights fundamentally disempowers them as political and economic agents (undermining their status as free and equal citizens), issues that will be elaborated in the next section.

The point here is that the terms of the education commodity proposition affect social relationships in schools. Testing-intensive determinations of educational value tend to create situations in which strategic relationships take precedence over communicative and collaborative ones (Habermas, 1984). Objectivity gives way to objectification, as students are caught up in the cost-benefit calculations of teachers and administrators. Instead of participating in relationships

based on reciprocity and mutual understanding (i.e., non-strategic and communicative), students find themselves having to navigate relationships that are the result of their placement into test-based bureaucratic categories (i.e., strategic and instrumental). These test-mediated educational relationships blur *the distinction between the students' needs and what school leaders and teachers are required to get students to do*; the former requires the establishment of a non-strategic communicative relationship, the latter requires taking up a strategic-instrumental relation, seeing only what the test says about how students should be treated in order to "improve the numbers." As Danziger (1990, p. 109) argues, testing practices in schools provide "a culturally acceptable rationale for the treatment of individuals by categories that bureaucratic structures demand." Or as Porter (1995, p. 77) says more broadly, "Numbers have often been an agency for acting on people, expressing power over them.... Numbers turn people into objects to be manipulated."

This transformation of the student from a unique individual into a standardized number—this process of testing-enabled objectification—is one means by which the education commodity proposition perpetuates injustice. As explored below and in later chapters, efficiency-oriented testing practices (which subsume the terms of the education commodity proposition) predispose students toward understanding themselves (and being understood by others) according to the categories they are put into as a result of numbers they receive on tests. When the tests lack objectivity this results in categories that are potentially arbitrary and misleading at best; at their worst these categories can create a kind of systematically distorted self-understanding (consider being miss-diagnosed with a learning disability). Moreover, the vast majority of test-based categories are *demonstrably inappropriate* as a source of self-understanding. The tests are not even based on a system of psychologically realistic concepts or theories, which is due to the

divorce of psychometrics from advances in the learning sciences (a point discussed again in the conclusion).

Moreover, and more importantly, students are powerless to change or even influence the system of test-based categories into which they are placed and which come to shape how they are treated and how they understand themselves. Put aside the fact that test bias and error commonly result in violations of students' rights to objective measurement (e.g., violating the first principle of just institutionalized measurement). The second and third principles of just institutionalized measurement emphasize the right of those affected by measures to have a voice in shaping measurement practices, both to assure their relevance and to secure possible benefits. Overriding these metrological rights results in testing practices that undermine the educational conditions necessary for students to understand themselves as free and equal while also creating inequities of opportunity and radically truncating the possibilities for self-actualization. This is the impact of measurement practices oriented toward abstract ideals of efficiency, seemingly untethered from the obligation to respect the rights of the individuals involved.

The dynamics of testing-intensive educational reform: between efficiency and justice

> Public statistics are able to describe social reality partly because they help define it. In the industrialized West...quantification has been part of a strategy of intervention, not merely of description.... As with the methods of natural science, the quantitative technologies used to investigate social and economic life work best if the world they aim to describe can be remade in their image.... To the

> extent that it [e.g., the measure] has become real, it provides the basis for a crucial kind of self-discipline... Measures of profitability—measures of achievement in general... provide legitimacy for administrative actions, in large part because they provide the standards against which people judge themselves.... Measures succeed by giving direction to the very activities that are being measured.
>
> -Theodore Porter (1995, pp. 43-45)

The focus of a theory of just educational measurement is not a static testing infrastructure in a static school system; the focus is the complex reality of the co-evolution of schooling and testing. Testing infrastructures that secure (or violate) certain metrological rights make it possible (or impossible) for educational institutions to make good on their commitments to social justice. The injustice of a testing infrastructure can be revealed through its effects on the educational institutions in which it is embedded—the test misshapes the school. At the same time, the injustice of a philosophy of education can be revealed through its effects on the technologies of testing it creates or adopts—the school misshapes the test. The causality implied here is not linear or unidirectional, but complex and dynamic. Both directions of analysis—that testing shapes schooling *and* schooling shapes testing—are needed in thinking through any given testing practice because both dynamics are always simultaneously in play.

This kind of co-determination or reciprocal constitution is a common theoretical trope in the social sciences, especially where what is being explained is an evolving institutional formation (Bowles & Gintis, 1986; Habermas, 1997). The classic statement of this view is "people create cultures and cultures create people" (e.g., people come to be who they are because of the culture they live in, while this culture itself is the result of the actions of the people who

live in it). The idea in this case is that school systems create testing infrastructures and testing infrastructures create school systems. This is not a paradox, but a way of thinking about social realities that requires some degree of comfort with non-linear causality and normative facts.

Importantly, this way of thinking reveals that some socio-cultural dynamics are prone to self-reinforcing directionality: authoritarian cultures create authoritarian individuals, who further contribute to the creation of a more authoritarian culture, which then creates even more staunchly authoritarian individuals, and so on in a kind of autocatalytic spiral (Bowles & Gintis, 1987). Analyzing the impacts of testing requires characterizing the complexity of this kind of self-reinforcing directionality as an important property in the dynamic between schooling and testing. By instituting a testing infrastructure, a school is changing its culture and practices; and these changes will tend to raise the relevance, valence and importance of the test, which will in turn create institutional conditions more conducive to that form of testing, and so on. Embracing the complexity of testing and schooling as co-constitutive allows for an investigation of one side of the dynamic (e.g., testing) for the traceable effects of the contradictions and unintended consequences stemming from the other side (e.g., schooling), and vice versa. So it is that testing infrastructures designed and implemented in the name of social justice can result in educational practices that are demonstrably unjust.

The dynamics of testing-intensive reform are complex, so they are grouped here under two broad headings: *efficiency-oriented* testing-intensive reforms and *justice-oriented* testing-intensive reforms.

Between efficiency and justice

There is a perennially recurring distinction between two tendencies in testing-intensive educational reforms. Both tendencies can often be found occurring in the same historical context and can be thought of as competing with each other in shaping the trajectory of the educational system. The terms have already slipped into the discussion before this point and are henceforth abbreviated as: *efficiency-oriented testing* and *justice-oriented testing.* These concepts are offered as ideal types and not as inductive generalizations or empirical classifications. They are intended to help bring clarity to the application of the theory of just educational measurement argued for here. As discussed above, this theory has a variety of moving parts, including the three principles of just institutionalized measurement and the three commitments of a just educational system, as well as a distinction between psychological and physical measurement, the education commodity proposition, and a theory about the co-constitutive dynamics of testing-intensive educational reform. The two ideal types are intended only to put a name on the most important recurring patterns in the application of the theory—marking two distinct constellations of insights that follow from the theory's conceptual architecture.

Efficiency-oriented testing is characterized by a recurring pattern in which the pursuit of objectivity is taken to extremes. Backed by the terms of the education commodity proposition, testing practices strive for objectivity but ignore the other metrological rights and thus establish educational institutions that are severely constrained in their ability to administer what social justice requires. Schools become places pursuing "the justice of efficiency" and display a tendency to institute forms of testing that further this ideal. As mentioned in Chapter 1, Rawls maintains that efficiency is an important institutional virtue, and that it is also a prerequisite for

administering justice, but efficiency is not synonymous with justice. This is because, simply put, *justice is an end-in-itself*, whereas *efficiency is a means-to-an-end*.

The pursuit of efficiency requires the prior establishment of institutional goals—efficiency is about the quality of the means used to accomplish a goal and not about the quality of the goal itself, which often remains unquestioned during its pursuit. Efficiency is an evaluative term, in the sense that it is applied with clear evaluative connotations: inefficiency=bad / efficiency=good. Yet efficiency is also *a morally neutral term*. This is a lesson brought home clearly by considering the ruthless efficiency with which the Nazis perpetrated genocide. Regardless of what the institutional goal is, it can be pursued more or less efficiently.

The pursuit of justice, on the other hand, is not defined relative to specific institutional goals; justice is, so to speak, a meta-goal for all the institutions of the basic structure. Achieving justice requires that all institutions strive for efficiency—justice requires some degree of efficiency—but optimizing efficiency does not necessarily lead to justice. Moreover, injustice itself is often inefficient. The *inefficiency of injustice* has shown itself in many attempts to improve efficiency at the expense of justice. For example, this has been the case where the key measures used for system surveillance have become increasingly error-prone due to indifference or subversion on the part of those using them (a point returned to in Chapter 5 in a discussion of cheating scandals in schools).

Where the tendency to efficiency-oriented testing dominates there is likely to be an educational system that receives its guiding social philosophy from sources beyond itself—typically from powerful stakeholders in the institutions that surround it and depend on its educational "outputs"— namely, economic institutions and the government. Efficiency-oriented testing is almost always undertaken in pursuit of goals set *for* the educational system by these

outside sources. These social goals, often institutionalized in the form of standards, transcend but include the general goal of increasing mere economic efficiency. This is a point that will be returned to in Chapters 4 and 5, where it will be shown that a variety of goals that have been set for the US educational system—from facilitating workforce restructuring to aiding national security—have initiated a variety of efficiency-oriented testing dynamics.

When efficiency is framed in terms of educational standards (which are then tied to tests) the result is a complex form of *social efficiency*, which goes beyond the terms of the education commodity proposition. This distinction between economic efficiency and social efficiency is critical (Lane, 1991, pp. 314-316): social efficiency occurs when the goals being pursued are outcomes in the social world (e.g., happiness; citizen virtues) as opposed to outcomes that are physical/productive (e.g., manufacturing more and faster cars, or designing more efficient computing) or financial (e.g., increasing profits). "Economic efficiency may actually frustrate social efficiency.... The efficiency norm that takes as its premise the maximization of profit does not serve the efficient pursuit of happiness.... Economic efficiency implies social *in*efficiency whenever it interferes with the optimization of other social goals with higher priority."

As discussed at the end of Chapter 2, Rawls argues that justice cannot be reduced to efficiency. It is also true that social efficiency cannot be reduced to economic efficiency, and so neither can the dynamics of efficiency-oriented testing be reduced to the terms of the education commodity proposition. In a school system with abundant resources (that does not need to prioritize economic efficiency), there will still be a need for efficiency-oriented testing of some kinds, as in, for example, program-evaluation studies or studies concerning the distribution of essential educational goods. However, in school systems where economic efficiency is a major concern, there are still cases where social efficiency trumps economic efficiency, as in the case

of special programs for the learning disabled, who from the perspective of sheer economic efficiency appear as a "drain" on the system. In these cases it is the acceptance of certain educational standards and policies that make the success of these programs part of the social efficiency of the school and thus expand the terms of the education commodity proposition.

Figure 2 shows the structure of efficiency-oriented testing, which subsumes the terms of the education commodity proposition (and its sole emphasis on economic efficiency). The parameters of what counts as social efficiency are determined by educational standards, which are directly related to what is tested. The standards define what matters, the tests measure learning in those terms, and so the testing results tell us the extent to which we are meeting the standards and how efficiently we are doing this.

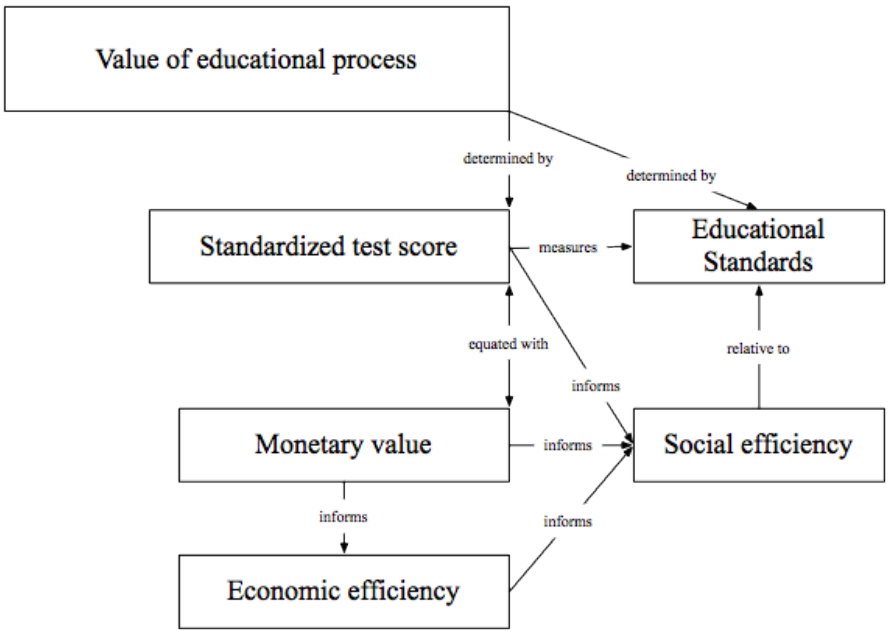


Figure 2: *efficiency-oriented testing*

As discussed above, the simplest manifestation of the education commodity proposition is when the value of an educational process is reduced to numerical units that can be monetized, something typically achieved through testing. When this approach to understanding the value of educational processes is subsumed within the broader pursuit of certain educational standards then the patterns that characterize efficiency-oriented testing begin to emerge. The education commodity proposition focuses on the economic efficiency of an educational process. Economic efficiency is in turn subsumed under a broader social efficiency that is the focus of efficiency-oriented testing. The education commodity proposition provides for considerations of economic efficiency but it leaves untouched the parameters that define the *social* efficiencies of the school or classroom: it allows for calculations of education-for-the-dollar but does not say what should officially count as valuable education. Efficiency-oriented testing sets in when economic efficiency is put in context by assigning a higher-order value to a set of standards, which establish what counts as valuable education.

Efficiency-oriented testing infrastructures are built to objectively measure only what matters in serving the overall social efficiency of the school system. At its most extreme, this leads to a situation in which educators (and others with a stake in the system) question the value of anything outside what can be measured through the current testing regime. Schools or teachers that pursue alternative goals are perceived as enemies of efficiency, and sometimes as perpetrators of injustice. If cost-benefit analysis is the catchphrase for the education commodity proposition, *quality control* and *continuous quality improvement* are the buzz words of efficiency-oriented testing. Here measures are used not just to assure economic efficiency, but to

monitor ongoing quality and to improve quality incrementally, through data-driven reforms.

Measurement-intensive quality-control processes have been widely instituted in manufacturing, business and government, all involving the simple idea of *repeated quality sampling* (Busch, 2011). The quality of any product constrains the economic efficiency with which it can be produced (it can be made only so cheaply before it devolves into something else). To maintain quality and thus regulate economic efficiency one must monitor the production line and institute a feedback loop between the quality of the outcome and the variability of relevant aspects of the production process. This is a common and necessary practice, especially in more complex production processes, such as those in the food and drug industries.

There are a common set of issues that characterize efficiency-oriented quality-control testing (Busch, 2011). Firstly, it is costly to test anything, and some testing practices necessarily damage or use up what is tested. You cannot test beer without drinking some of it, nor test the quality of fuel without burning it. Moreover, testing procedures have their own intrinsic cost, including measurement equipment, project management, and intermittent production and personnel overhauls—the so called "costs of surveillance" (Busch, 2011; Bowles & Gintis, 1986). But most important is the power of measurement-intensive practices to control and transform the social processes being regulated. This is what Busch (2011, pp. 28-32) focuses on in his far-reaching sociological analysis of standards and their related measurement practices, in this case explicitly looking at educational testing:

> [There is an] intimate connection between standards and power... in our modern world standards are arguably the most important manifestation of power relations [considered by many as "soft laws"].... But this is not to suggest that standards

> have the kind of power we associate with a tyrant... To the contrary, the power of standards lies in their becoming barely noticed. [Standards and their related measures] *set the rules that others must follow, or set the range of categories from which they may choose*... Standards display anonymous power. Even if we know who established them, standards take on a life of their own that extends beyond the authorities in both time and space.... As C. Wright Mills showed 50 years ago, those who fail exams see public issues (the small chance of passing, the legitimacy of the examination system itself) as personal troubles (the failure to study sufficiently hard, to memorize the necessary texts, etc.). The highly standardized exams were "naturalized." They were not subject to challenge but were seen as challenging to those who took them.... In short, the power of established standards [and related measures] is that they structure our expectations, because standards, like the world of nature, are seemingly "supposed" to be the way they are. What could be more powerful than something that is revealed as no less than a part of the natural world itself?

The liabilities that accompany the institutionalization of measurement-intensive quality-control standards also accompany efficiency-oriented testing initiatives in schools. As discussed in Chapter 5, standards-based reform movements in education end up spending a relatively large percentage of their budgets on the *surveillance* necessary to perform quality-control monitoring. This efficiency-oriented quality-control testing has an impact on the culture and policies in the schools. Here again is the problem of *damage done to the product through repeated quality sampling*—students cannot be tested without being impacted, even if the impact is only on the

use of their time. Some tests actually damage students—like fuel burned during a test of its quality. But in the case of fuel, a small sample can be taken as representative of a large quantity. In current educational policy, however, *every student is tested*. It is true that generalizations are made by aggregating and disaggregating different student subgroups, and quasi-experimental designs do characterize program-evaluation studies—but nevertheless for the purposes of most standards-based reforms, every single student is tested as a part of quality-control surveillance. This means that the impacts of repeated testing (e.g., time spent in testing and test-prep; test anxiety) are guaranteed to be widely distributed. As in some industrial contexts, there is a potential with this kind of quality-testing that the damage done to the product (or the exorbitant expense of) attempting to measure its quality outweighs the benefits sought from measurement in the first place.

Standards-based concerns about quality have overseen the institutionalization of testing infrastructures that "*set the range of categories from which all may choose*... [through a kind of] anonymous power" (*ibid*.). Efficiency-oriented testing involves specific definitions of social efficiency. These are provided by educational standards that are institutionalized in classroom practice, curricula, and school management, and then monitored through testing, which provides an index useful for both economic and social efficiency. The standards define the vision of social efficiency toward which the school strives and in terms of which it is determined to be succeeding or failing. Once in place the standards-based norms of school practice become the taken-for-granted backdrop against which individuals live out their educational biographies.

Standards regulate both testing and the terms of economic efficiency, but standards themselves must derive their legitimacy from a broader social philosophy. Educational standards

do not justify themselves—they are dependent on a supporting social philosophy or philosophy of education. This provides an opening for *justice-oriented testing practices*.

Justice-oriented testing

Rawls would argue that standards and policies ought to be backed by a theory of justice because they function as shared terms of cooperation (defining the social goals pursued by everyone in the system). Standards suitable for a just educational system must be built so that they are the subject of broad consent. Ideally, the range of application of the education commodity proposition is determined by a form of social efficiency that is determined by a theory of justice—such an arrangement is likely to involve *justice-oriented testing* (Figure 3).

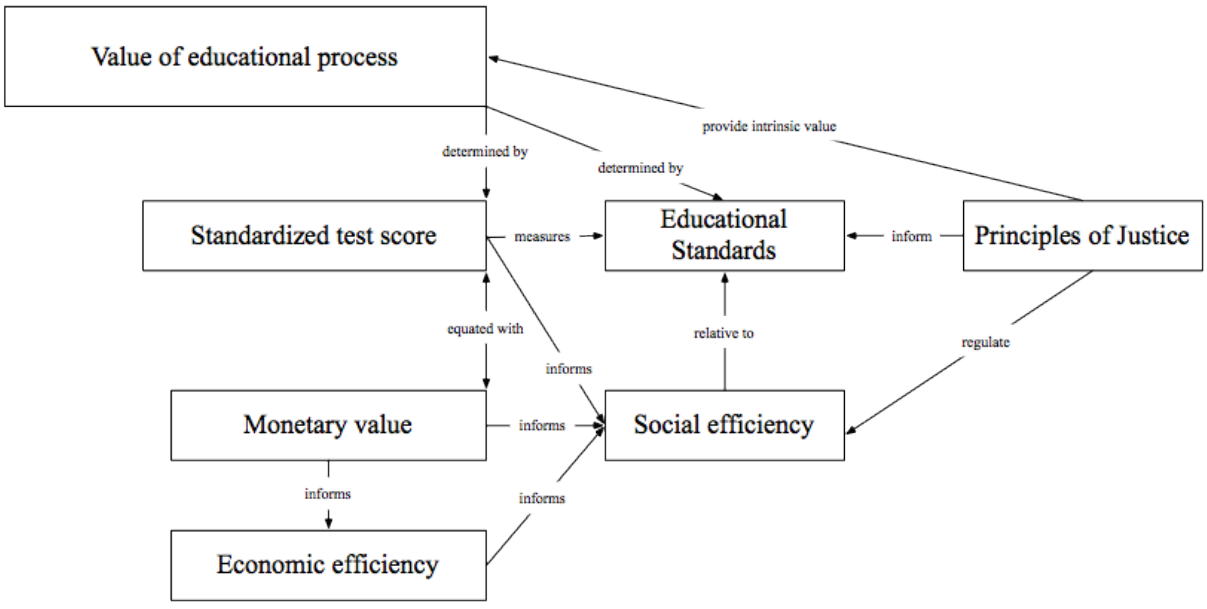


Figure 3: *justice-oriented testing*

In a just society, the educational system provides an essential entitlement, a basic right—it administers the fair allotment of educational primary goods. Justice-oriented testing is characterized by a recurring pattern in which the full set of students' metrological rights are considered as a means for actualizing the commitments of the educational system to social justice. Justice-oriented testing infrastructures are built to assure that objectivity is secured for all, but not at the expense of relevance or the possibility of directly benefiting students. This kind of testing infrastructure creates (and is created by) an educational system preoccupied with providing for the full range of educational primary goods, contributing to a system of background institutions that promote fair equality of opportunity and facilitate the self-actualization of all students.

Needless to say, there has never been a pure instance of justice-oriented testing, just as there has never been a pure, ideal instance of efficiency-oriented testing. But indications of this dynamic have been in evidence since the first testing infrastructures were built in the early decades of the twentieth century. Testing infrastructures have been built to shoulder the burden of accomplishing a variety of functions that are a necessary part of an educational system committed to social justice, including administering a kind of pure procedural justice in the allocation of opportunity, assuring the equitable distribution of educational primary goods and identifying the unique learning needs of each student.

Justice-oriented testing is based on a fundamentally different way of thinking about educational reform than its efficiency-oriented counterpart. The goals to be met by justice-oriented testing are not given to the educational system and classroom, but emerge organically from within it. Rawls's lessons about education imply that social justice should be conceived as

an intrinsic part of what educational systems are built for, unlike the wider variety of goals so often foisted upon education by its bordering institutions.

At its most extreme, justice-oriented testing reverses the typical dynamic between the educational system and the other institutions of the basic structure, positioning the goals of the educational system (e.g., it social justice mission) as superordinate to both economic and governmental goals. According to this view, the dynamics of education reform must expand outward from the schools to touch every sector of society because many of the nonnegotiable commitments of the educational system cannot be met through the reform of schools alone. This is an idea that will be discussed at length in the Conclusion, where the radical implications of designing justice-oriented educational systems and testing infrastructures are elaborated. The point to make here is that justice-oriented testing replaces the notion that "efficiency *is* justice" with the notion that "efficiency *serves* justice"—thus positioning justice as the dominant institutional virtue and rearranging the commitments and relationships of the other institutions of the basic structure accordingly.

The distinction between efficiency and justice is at the heart of the theory of just educational measurement argued for here. This distinction reflects Habermas's (1984; 1987) distinction between "strategic-rationality" and "communicative-reason" and Rawls's (1996) idea of "the Rational vs. the Reasonable," which both retrofit Kant's (1785) distinction between categorical imperatives and hypothetical imperatives. They all maintain that disinterested and self-interested calculation do not exhaust the possibilities of human reason. As Rawls (1996, p. 48-54) explains, the distinction is often made in everyday speech: "We say: Their proposal was perfectly rational, given their strong bargaining position, but it was highly unreasonable, even outrageous.

> Reasonable persons... desire for its own sake a social world in which they, as free and equal, can cooperate with others on terms they all can accept. They insist that reciprocity should hold within that world so that each benefits along with others. By contrast, people are unreasonable... when they plan to engage in cooperative schemes but are unwilling to honor, or even to propose, except as necessary for public pretense, any general principles or standards for specifying fair terms of cooperation. The rational is a distinct idea from the reasonable and applies to a single unified agent (either an individual or corporate person) with the powers of judgment and deliberation in seeking ends and interests peculiarly its own. The rational applies to the choice of means, in which case it is guided by such familiar principles as: to adopt the most efficient means to ends, or to select the more probable alternative, other things equal.... Knowing that people are rational we do not know the ends they will pursue, only that they will pursue them intelligently. Knowing people are reasonable we know they are willing to govern their conduct by a principle from which they and others can reason in common; and reasonable people take into account the consequences of their actions on others' well being.

With the definitions set in this way, the view expressed here is that efficiency-oriented testing (and the terms of the education commodity proposition) are perfectly rational, but they can often be entirely unreasonable. This is another way of saying what was said above: these forms of testing-enabled decision-making are *true but partial*. The pursuit of rational efficiency does not require the institutionalization of fair terms of cooperation, it only requires the

intelligent (evidence-based, data-driven) pursuit of goals. Neither the goal nor the means used to attain it need to be acceptable to everyone involved. From the perspective of rationality, both democratic decision-making and "stakeholder buy-in"—any form of reciprocity, really—are understood as constraints on system efficiency; too little buy-in can reduce efficiency, but too much democracy is also inefficient (Habermas, 1987). Generally, rationality demands as little democracy as possible to maintain legitimacy, whereas being reasonable requires some form of democratic decision-making as a nonnegotiable aspect of institutional culture.

Of course, depending on what goals are set, efficiency-oriented testing can contribute to educational institutions that meet some commitments to social justice, such as providing for certain educational primary goods. Yet making good on these commitments to justice is an *unintended outcome* of increased efficiency, even if it is a welcome one. If the broader goal could be met in a more efficient way without making good on these commitments, then that would be the preferable trajectory of reform. This should bring to mind Rawls's criticism of the utilitarian "liberalisms of happiness" discussed in Chapter 1. This family of political views takes *certain basic rights* as *negotiable* in the strategic pursuit of gains in *aggregate happiness*. Likewise, efficiency-oriented testing infrastructures create (and are created by) educational institutions in which the basic rights of students are negotiable in *the strategic pursuit of system-level goals*. Efficiency-oriented testing is thus often rational but unreasonable—it tends to establish a system of cooperation that "works" for a small number of decision-makers but that could not meet with the reasonable consent of all those involved.

Putting the theory to work: making sense of the history of testing

This chapter has introduced the components of a theory of just educational measurement: 1) the three principles of just institutionalized measurement; 2) the three commitments of a just educational system; 3) an understanding of the (dis)analogies between psychological and physical measurement; 4) the "logic" of the terms of the education commodity proposition; and 5) the dynamics of testing-intensive reform, best summarized in terms of the difference between efficiency-oriented testing and justice-oriented testing.

The next two Chapters focus on putting this theory of just educational measurement to use as an analytical tool. The goal of these discussions is not to expand upon or improve the historical record, rather the goal is to use established historical accounts as grist for the theoretical mill—putting key "facts" into the framework to see how it handles them. For those unfamiliar with the history of testing, the rest of this work should serve as an eye-opening introduction to a fascinating and important area of sociological and historical scholarship. For those more familiar with the history of testing, it should be a provocative and insightful rereading of some of the key episodes. The theory will have proven itself if it is able to consistently make sense of those considered judgments deemed uncontroversial while also providing greater clarity on more ambiguous and problematic ethical issues.

4: Social justice and the origins of educational measurement

> Test scores influenced the behavior of professionals and the self-concept of the children.... [The scores] seemed to prove that the social order was close to a meritocracy since the fittest seemed mostly on top. They helped to fix on mass institutions of education, civil service, and business narrow standards of what constituted ability. All this was no malevolent plot.... psychologists were men trying to make democracy work efficiently in what they believed was a great cause. They even had their moments of utopian dreaming of a smoothly running, conflict-free society where talent rose and ruled benignly. But the effects of their technology of objective discrimination needs to be assessed as well as their intentions.
>
> -David Tyack (1974, p. 204)

> Mental testing flourished because it helped the American schools to be both a comprehensive and a differentiated institution; the tests squared American ideals of equality of opportunity with a social structure that resembled a pyramid.
>
> -James Reed (1990, p.77)

For as long as there have been schools there have been tests of one kind or another. In fact, even before schooling was widespread, tests played a critical role in the intergenerational transmission of skills and culture (Duncan, 1984). A test can be as simple as checking in with a

child to see if he is able to do what he has just been shown (e.g., "tie the knot I just showed you") or as complex as the forms of scientifically standardized mass-administered practices that are the overall focus of this study. But at its most basic, some kind of testing is an essential part of all educational encounters because there is no other way to know if the learning that was intended to take place has indeed taken place. There are accounts of testing practices in ancient Greece and Egypt, both for artisans and scholar-priests, but it is hard to imagine farmers and merchants transmitting skills without them. The system of Mandarin examinations dates back even further, beginning in the earliest history of Chinese civilization; these are probably the first mass-administered high-stakes examinations for the purpose of selecting individuals for bureaucratic positions. For millennia, tests were an essential part of religious education throughout the world, but it was in the monasteries of Medieval Europe (and then in the universities they initiated) that a system of examinations and credentialing began to take shape that would eventually spread to American soil and take up root in colonial era schools (Cremin, 1970). The remnants of this system are still obvious (especially in the Ivy League), from the oral exams that accompany the award of a doctorate (itself a credential of medieval vintage) to the gowns worn at graduation and the Latin text on the diplomas handed out.

Testing in some form has always been around, but standardized testing infrastructures as they exist today are an invention of twentieth century America. Standardized tests are fundamentally different than their diverse and ubiquitous progenitors. Aside from obvious differences in format, administrative procedures, and content, the heart of the difference resides in the relationships established through the testing practices themselves. Pre-standardized testing practices were not conducted under the pretense of objective and scientific measurement. Most were conducted as communicative exchanges and socio-cultural practices embodying the values

of the teacher, who stood as a kind of mediator between the student and the broader culture. Make no mistake, traditional testing was fraught with bias, filled with seemingly irrelevant content, and, notoriously, were attended by harsh (often physical) punishments. Yet, in many cases, testing was conceived as part of the teacher's responsibility to their students, a practice almost entirely at the teacher's discretion that emerged from within the educational dynamics of the student-teacher relationship. By contrast, standardized tests are not conceived as part of the culturally mediated conversation between the teacher and the student. They are intended for use as scientific measurement instruments, ideally devoid of culture-specific values, and best administered in impersonal contexts that resemble more the relation between scientist and subject than between teacher and student. Moreover, standardized tests are accompanied by a range of theoretical commitments that guide their design and implementation, from scientific theories about the constructs being measured to social philosophies determining how the results should be used. Finally, these tests are deployed on a scale and for purposes that would have been inconceivable to educators as late as the mid-nineteenth century. Standardized testing infrastructures have been predominately used to enable the "scientific management" of massive school systems. This has involved in millions of students be given tests designed and administered by experts working for organizations outside the schools in order to assure objective systems-level data is be gathered for administrators, politicians, researchers, and reformers. No longer a practice emerging from the relationship between teacher and student, standardized tests are imposed upon this relationship by authorities who hold the culture of the school at arm's length.

It is interesting to note that even the system of grading students from A to F was not adopted as a universal practice until the 1890s. This grading system and its related numerical

values (the GPA)—now a seemingly ahistorical staple of schools throughout the world—was invented at Mount Holyoke College in 1887, giving standard form to grading practices that had been in flux for nearly a century (Durm, 1993). This particular approach was inspired by those being used in the newly expanding factory system, where both products and workers were graded similarly. Its widespread use and eventual near-universal adoption is also a testament to the increasing importance of standardization in industrial production in the US during the last decades of the nineteenth century.

Congress established the Office of Weights and Measures in 1836 to assure the national standardization of critical instruments and processes, yet as late as 1886 its newly appointed head—the philosopher, Charles S. Peirce—would complain that "the Office of Weights and Measures is a very slight affair," lacking all necessary equipment for carrying out its intended tasks (quoted in Dupree, 1957, p. 271). But this was soon to change as industrialists took matters into their own hands, working to persuade Congress to establish the National Bureau of Standards in 1902, which was to subsume the Office of Weights and Measures and usher in an era during which the standardization of nearly all aspects of industry was carried out at an astonishing pace.

As one of the central architects of the Bureau testified before Congress: "[There is] a moral aspect of this question; that of recognition by the government of an absolute standard, to which fidelity in all relations of life affected by that standard are required. We are victims... of a lack of comprehension of the binding sanction of accuracy in every relation of life.... Nothing can dignify this government more than to be the patron of and the establisher of absolutely correct scientific standards...." (quoted in Nobel, 1977, p. 75).

In this quote there are echoes of the ethical discourse that surrounded the creation and dissemination of the metric system, as discussed in Chapter 1. Beyond the obvious benefits for industrial and economic efficiency that would result, the first principle of just institutionalized measurement—the right to objective measurement—is invoked. Ethical motives are invoked for instituting governmental regulation of weights and measures. The idea that government is dignified through securing "the binding sanction of accuracy in every relation of life" is based on a realization that measurement infrastructures are basic structures, which implicate everyone in society and are thus subject to evaluation in terms of social justice. Thus the drive toward the standardization of measurement instruments gained widespread support and was tied into a broader ethical zeal for *the justice of efficiency*, discussed in greater detail below.

The goal of establishing "national standards for all aspects of economic life" was quickly (and logically) expanded to include people. As one leader of the newly emergent movement of scientific labor management put it, "the extension of the principles of standardization to the human element in production is the most important growing field of activity" (quoted in Nobel, 1977, pp. 83). This was the context into which the psychometric work of Alfred Binet was imported from France.

Testing in the name of the least well-off: Binet's vision of testing and justice

> If the impression takes root that these tests really measure intelligence, that they constitute a sort of last judgment on the child's capacity, that they reveal "scientifically" his predestined ability, then it would be a thousand times better if all the intelligence testers and all their questionnaires were sunk without warning

> into the Sargasso Sea. One has only to read around in the literature...to see how easily the intelligence test can be turned into an engine of cruelty, how easily in the hands of blundering or prejudiced men it could turn into a method of stamping a permanent sense of inferiority upon the soul of a child.
>
> -Walter Lippmann (1922)

Binet is justly famous as the inventor of the IQ test, which was an ingenious solution to a difficult problem put to him by the French minister of public education in 1904: devising a means for identifying students who could not be placed in normal classrooms because they required special education. (While accounts of Binet's invention and its subsequent mutations can be found in almost every psychology text book, the one offered here draws from several sources: Brown, 1992; Gould, 1996; Lagemann, 2000; Sokal, 1990.) His approach to this problem would change the face of psychology and education, with ramifications affecting an incredible array of institutions and cultural practices. Importantly, Binet's work came to be used and understood in ways that Binet himself would have strongly opposed, so it is worth looking more closely at the original procedures and ideas surrounding the birth of standardized testing. As will become clear, Binet intended his instrument for use as a part of justice-oriented testing practices, but his intention was lost in the enthusiasm for efficiency that dominated the contexts of its American importation.

Before Binet's invention there were a wide array of competing approaches to psychological measurement, most of which involved commitments to a kind of faculty psychology tracing its lineage to phrenology and other physicalistic means of determining individual differences. As early as the 1880s the pioneering American psychologist James Cattell

(who coined the term "mental test" in 1889) began importing the instruments and techniques of Francis Galton's "anthropometry," which consisted of standardized physical apparatuses for detecting minute differences of sensory and motor capabilities, such as "reaction time to sound" and "least noticeable difference in weight" (Sokal, 1990). The results of the various physical tests were often taken as a proxy for a variety of psychological faculties, such as perceptiveness and perseverance. Cattell was instrumental in putting together an exposition of the new science at the 1893 World Fair in Chicago. Thousands of individuals were tested, greatly increasing a growing public fascination with the standardized and scientific measurement of minds. But the popularity of this approach would be short lived, due in part to growing awareness of the work being conducted by Binet. Emerging criticisms focused on the limits of physiological measures as indices of meaningful psychological differences. A growing desire spread through the profession for "giving tests as psychological a character as possible" (James Mark Baldwin, quoted in Sokal, 1990, p. 35).

Binet's tests offered just that. They made no use of complex physical apparatuses and involved linguistically mediated tasks that clearly elicited the so-called "higher mental processes." When compared to the 'mad-scientist' laboratory of anthropometric testing instruments, Binet's tests appeared much more similar to the examinations given in schools for centuries to determine the knowledge and skills possessed by students. Yet Binet's tests were fundamentally different from traditional forms of academic evaluation. These differences revolve around the requirements of standardization and objectivity.

For one, he was not interested in "learned skills" such as reading and mathematics, nor was he interested in the knowledge associated with traditional academic subject matter. Instead he aimed to bring together a large series of seemingly everyday tasks, such as counting coins or

determining which of four female faces were "prettier." The tasks were thought to get at more general processes of reasoning. The idea was that mixing together a wide range of tasks would allow for an inference to the child's general ability.

The various tasks were administered one-on-one by trained examiners in a sequence scaled by their order of difficulty. Each order of difficulty was assigned an age level, defined as the youngest age at which a child of normal intelligence should be able to complete the tasks. The child began with tasks for the youngest age and proceeded up the scale until they could no longer get them right. During the first decade of its use there was a variety of ways in which the results were quantified, but researchers eventually settled on a common method. The child's "mental age" was indicated by the last task in the age-graded scale they could complete. Their "general intellectual level" was then determined by dividing this test-determined mental age by their actual chronological age (multiplying the result by 100 to eliminate the decimal point) and thus the *intelligence quotient*, or IQ, was invented.

Binet was interested in the degree of the discrepancy between a child's mental age and his or her actual age. Knowing that a child's mental age was greatly behind his or her chronological age allowed them to be identified as in need of special educational accommodations. Indeed, this was the only reason the test was invented. Binet consistently stressed the pragmatic and empirical nature of the scale and "consistently declined to award any theoretical interpretation to his scale of intelligence.... [He also] declined to define or speculate upon the meaning of the score he assigned to each child" (Gould, 1996, p. 180). He argued that:

> intelligence is too complex to capture with a single number.... The scale properly speaking, does not permit the measure of intelligence, because intellectual

> qualities are not a single scalable thing like height.... We feel it is necessary to insist on this fact, because later, for the sake of simplicity of statement, we will speak of a child of 8 years having the intelligence of a child of 7 or 9; these expressions, if accepted arbitrarily, may give place to illusions.
> (Binet, quoted in Gould, 1996, p. 181)

Even more important than Binet's theoretical reservations about the interpretation of scores awarded by his test were his ethical and pedagogical concerns about its possible and preferable uses. In fact, Binet's ideas about the use of his instrument touch on all three principles of just institutionalized measurement, and thus represent a *profound alternative* in educational measurement, the *road not taken*.

Binet understood the essential need for objectivity in an instrument designed to serve such critical institutional purposes and built the test accordingly. But he worried about the potential harm that could be done when the scores took on an institutional life of their own. Binet pleaded passionately on behalf of the learning-disabled and protested against the use of his tests in ways that stigmatized the child:

> If we do nothing, if we don't intervene actively and usefully, he [the learning-disabled child] will continue to lose time.... and will finally become discouraged. The situation is very serious for him, and since his is not an exceptional case (since children with defective comprehension are legion), we might say that it is a serious question for all of us and for all of society.... [Shame on those] teachers who are not interested in students who lack intelligence. They have neither

> sympathy nor respect for them, and their intemperate language leads them to say such things in their presence as 'This is a child who will never amount to anything... he is poorly endowed... he is not intelligent at all.' How often have I heard these imprudent words... Some recent thinkers seem to have given their moral support to these deplorable verdicts by affirming that an individual's intelligence is a fixed quantity, a quantity that cannot be increased. We must protest and react against this brutal pessimism; we must try to demonstrate that it is founded upon nothing. (Binet, 1909, p.100-101)

Binet believed the tests should be used *only* as a means for helping the least well-off children in ways that were most relevant and beneficial to them. He even developed and implemented a program of "mental orthopedics" intended to supplement the use of the tests and aid children identified as needing special attention and guidance. Gould best summarizes Binet's "three cardinal principles for the use of his tests... all of which were later disregarded by the American hereditarians who translated his scale into written form as a routine device for testing all children:

1) The scores are a practical [objective] device; they do not buttress any theory of intellect. They do not define anything innate or permanent. We may not designate what they measure as "intelligence" or any other reified entity.
2) The scale is a rough, empirical guide for identifying mildly retarded and learning-disabled children who need special help. It is not a device for ranking normal children.

3) Whatever the cause of difficulty in children identified for help, emphasis shall be placed upon improvement through special training. Low scores shall not be used to mark children as innately incapable. (Gould, 1996, p. 185)

These were the principles intended to guide the use of the first scientifically refined standardized tests. They are fully congruent with the principles of just institutionalized measurement, positing each child's right be objectively measured in ways that are both relevant and beneficial. IQ tests implemented in schools according to these principles would be insulated from co-optation as part of efficiency-oriented testing practices, such as those discussed immediately below, where whole student bodies were ranked and sorted so that resources could be funneled away from the least well-off and towards those with greater "innate" abilities. It bears repeating: *this is the opposite of the test's intended use*. If Binet had had his way his tests would have been instruments used *only* to identify and help the least well-off, *period*. Moreover, having only this pragmatic use, with no accompanied theoretical meaning, the test score would not have served as an enduring label for the child, being *best forgotten by child and teacher alike*.

This last point—that the test score must not become a permanent label—is directly related to discussions in Chapter 3. Recall the potential for the terms of the education commodity proposition and efficiency-oriented testing to obscure how students understand themselves and are understood by others. The benefits of objectivity and quantification (and as discussed below, the benefits of mass administration and efficiency) must be weighed against the direct effect of testing on the social relations constituting the educational process. Testing practices, like other forms of measurement, set the terms of mutual understanding and facilitate coordinated

interpersonal activities. The meaning of the test for the student, teacher, and administrator conditions their relationship; it establishes a shared sense of "what is the case." Because of this inevitability stemming from the nature of testing, tests should be designed and used in ways that assure they do not create mutual understandings that are systematically distorted by the meaning of the test—as when, for example, the test scores are understood as predominately markers of "innate" or "inherited" differences.

Gould clarifies the issue in terms that reflect the difference between justice-oriented and efficiency-oriented testing:

> The differences between strict hereditarians and their opponents is not, as some caricatures suggest, the belief that a child's performance is all inborn or all a function environment and learning. I doubt that even the most committed antihereditarians have ever denied the existence of innate variation among children. The differences are more a matter of social policy and educational practice. Hereditarians view their measures of intelligence as markers of permanent, inborn limits. Children, so labeled, should be sorted, trained according to their inheritance and channeled into professions appropriate for their biology. *Mental testing becomes a theory of limits*. Antihereditarians, like Binet, test in order to identify and help. Without denying the evident fact that not all children, whatever their training, will enter the company of Newton and Einstein, they emphasize the power of creative education to increase the achievements of all children, often in extensive and unanticipated ways. *Mental testing becomes a*

> *theory for enhancing potential through proper education*. (Gould, 1996, p.183; emphasis added)

It is one of the great ironies in the history of testing that the man who invented the first and most widely used standardized test understood the social justice implications of his invention and articulated a vision to assure its appropriate use, only to be completely ignored by his most enthusiastic and ambitious followers. The individuals who so drastically repurposed Binet's invention, all of them Americans, were convinced of an extremely consequential theoretical commitment, namely that what the IQ test measured was a fixed, inherited trait. Their idea was that a person's intelligence is an inalterable inherited property of the mind that is best thought of as akin to a strictly biological trait, such as height, and that this trait could be measured objectively by an IQ test. This idea, which has supporters to this day (Herrnstein & Murray, 1994), was invented by American psychologists during a very specific historical epoch. Wedding this idea to Binet's invention allowed for a much wider range of ostensibly valid theoretical and institutional applications and turned what was potentially an instrument of justice into an instrument of injustice. The idea that IQ tests measured an innate ability would have a massive impact on the shape of American education during the first half of the twentieth century, as the proliferation of scientific racism backed advances in the mass administration and institutionalization of standardized testing. These developments in institutionalized measurement would, to take a phrase from Condorcet (1785), "make nature herself an accomplice in the crime of political inequality."

Social justice and the IQ testing movement in America

The rest of this chapter looks at the consequences of adopting Binet's invention divorced from his views of how it should be used to serve social justice. The discussion deals with a variety of historical episodes, which exemplify the use of IQ tests in the American public schools during the first decades of the 20th century. This section here outlines the central issues that will be thematized throughout this chapter (and, in a slightly different form, in Chapter 5). It will be clear that these themes are directly related to the theory of just educational measurement built over the course of the preceding chapters.

The primary social justice issue implicated in the early IQ testing movement was the profound *lack of objectivity* that characterized its major instruments and practices — a blatant violation of the first principle of just institutionalized measurement, and one that would set off a cascade of injustices throughout the educational system. For decades this lack of objectivity has been a common theme in critical histories of the early IQ testing movement, which have located the movement's lack of objectivity in the cultural biases of early IQ tests, thought to stem from the overt racism of the leaders of the testing movement. Many of these criticisms are valid, and links between Eugenics and the testing movement are discussed below. However, when it comes to the early IQ testing movement, issues concerning objectivity are actually more basic and require no *ad hominem* arguments.[6]

[6] Make no mistake: racism was a major problem throughout the American educational system during this period (and remains a major problem today). The arguments offered here are not intended to downplay this aspect of the historical context. Instead, the goal is to identify, at the level of the testing practices themselves, the illicit epistemological moves that allowed for the perpetuation of racist ideas by creating the *illusion* that they were backed by scientific theories and objective measures. When the early IQ testing movement is simply dismissed as racist, without further analysis, we have gained no insight into the mechanisms by which testing was made an 'accomplice in the crime of political inequality.' Moreover, we miss the fact than many of these mechanisms are still in play as aspects of contemporary testing infrastructures. This is discussed further in the section Eugenics below.

Recall the definition of objectivity offered in the Excursus that concluded Chapter 1. There it was shown that objectivity is a combination of reliability and validity — a psychological measure is objective when it represents what it is claimed to represent (validity), and when it does so in an accurate, consistent, and unbiased manner (reliability). Measures that are characterized by this kind of objectivity are a prerequisite for social justice in a wide range of contexts. IQ testing as practiced in the early decades of 20th century lacked objectivity on both counts, for the tests were neither valid nor reliable.

Their lack of reliability is easy to establish. As will be discussed in later sections, this reliability deficit was due to a simple lack of attention to item design and a failure to standardize administrative procedures and scoring. These two problems were compounded by the presence of hundreds of different tests on the market, and yet no generally accepted standards for reliability. This diversity of test on the market, in turn, was then amplified by the shoddy statistical procedures used in the interpretation of test results. Viewed by contemporary psychometric standards, these tests were deeply flawed; they lacked even the most basic qualities to be considered accurate, consistent, and unbiased. This fact alone should nullify whatever claims were being made about their results, whether about "innate intelligence" or anything else. Moreover, the tests' lack of reliability is also enough to undermine the legitimacy of using them for tracking students or making any other high-stake decisions.

Even if, for the sake of argument, we stipulate that in some schools the tests did *not* lack reliability — perhaps in schools that meticulously used the Stanford-Binet — the tests still lacked both construct and ecological validity. It is with these concerns over the validity of early IQ tests that things become both more interesting and more problematic. As has been made clear, Binet had a specific set of ideas about what his tests measured and likewise identified a particular

set of institutional circumstances in which they were to be used. In his hands the tests were arguably characterized by high levels of construct validity (it was measuring what he claimed it was measuring) and ecological validity (it was fit for serving the institutional uses to which it was being put). Unfortunately, the American adaptation of Binet's invention jettisoned his ideas about what it measured and how it should be used. The IQ test became not only a means for determining the level of students' genetically inherited intelligence, but was also used to track students into capability groups that permanently impacted their academic trajectory (i.e., not remedial aid for some, but wholesale academic segregation based on "innate intelligence").

Construct validity should be discussed first, because the institutionalized use of a test follows from what the test purports to measure. The leaders of the early IQ testing movement claimed that IQ tests measured intelligence, understood as a heritable, fixed trait that remained relatively unchanged over the lifespan. This form of intelligence was taken to be the dominant characteristic of an individual's mind, a claim that had (and has) no basis in reality. Put aside the fact that during the early decades of the 20th century there was no understanding of the structure of DNA, much less any robust scientific theories about how intelligence might be transmitted genetically (Kevles, 1998). And likewise, put aside that the very notion of a singular unified intelligence being the dominant characteristic of psychological life was (and is) an untenable psychological theory, dismissed even then by most psychologists who were not directly tied to the testing movement (Gardner, 2011; Gould, 1996). Indeed, even *if* we do stipulate that there is such a thing as a genetically heritable and unchangeable intelligence that determines the overall character of the mind, it is *still* the case that IQ tests do not come close to measuring anything like this.

The most basic problem has already been discussed: the tests lacked reliability. They were not accurately and consistently measuring anything, so all claims made about what the tests revealed or "proved" were specious. But, for the sake of argument, let us imagine some fictional school and test wherein reliability was not a problem. In this case it is the content and structure of the tests themselves that undermine the claims made about them. The tests consisted of a series of tasks ranging over a variety of "everyday" problems, mediated by language, and often, especially in the case of the earliest IQ tests, specific aspects of the dominant culture. Clearly, these were *not a direct measure of intelligence*, let alone some form of intelligence that is heritable and that constitutes an unchangeable dominant characteristic of an individual's mind.

For contrast take a simple vocabulary test, consisting of a list of words that the student is asked to define and use in a sentence. *This is a direct measure* of a student's academic vocabulary, albeit a fairly limited psychological property. Even so, no inference needs to be made concerning some indirectly measured construct for which the test serves as a proxy. At best, IQ tests can claim to be indirect measures of intelligence, but even then it is hard to justify this claim based on the content and structure of the test (e.g., Why these items? Why isn't the ability to hold a conversation, interpret a poem, or build a tree house a better measure of intelligence?).

Moreover, even if we stipulate that the test is a proxy for something like intelligence, there are still two problems. We need some other index of intelligence that can be used to validate the claim that the IQ test is a good proxy. And, more importantly, we need some other set of reasons for believing that what the tests measure is both heritable and constitutes an unchangeable dominant characteristic of an individual's mind. Keep in mind that neither of these exists — neither then nor now. Thus, early IQ tests lacked construct validity due to a series of

illicit inferences: the first being that any inferences can be made at all based on an unreliable test; second, that the test (even if it were reliable) serves as a proxy for intelligence; and third, that this form of intelligence is heritable and unchangeable and constitutes a dominant characteristic of the mind.

This leads us directly to the second set of social justice issues implicated in the early use of IQ testing in American schools: their use in tracking students for the sake of efficiency into capability groupings. This issue has two facets. The first falls under the heading of ecological validity, and has already been mentioned. That is, given what the tests (actually) measure and how well they do so, have they been institutionalized appropriately? The second issue concerns the nature of the academic tracking system itself and how it impacted the distribution of educational goods.

The first issue is clear. One of the cardinal rules of psychometrics is that when a test lacks both reliability and construct validity it should not be used to make any high-stakes decisions (AERA, 1999). The early IQ testing movement clearly violated this rule. The tests were institutionalized *as if* they measured something they did not (i.e., innate intelligence), and *as if* they measured it well (i.e., accurately, consistently, and without bias). Note that Binet did not make this same mistake and was careful to warn of the potential misuse of his invention, chiefly the possibility of it being institutionalized in ways would put permanent labels on children, and especially ones marking them as "deficient."

Of course, that is exactly what happened when the IQ testing craze swept through the American educational system. On the back of a broad social movement for economic efficiency, and in the face of a rapidly and massively expanding public education system, IQ tests were used to create a particular kind of tracking system within schools. Because of the (erroneous) beliefs

about what IQ tests measured they were used as part of a tracking system that reified test-based individual differences as innate characteristics. This unacceptably distorted the distribution of educational goods. The capability groups—understood as based on innate differences in genetic endowment—were made *far* too rigid and deterministic, and thus undermined all of the commitments of a just educational system outlined in Chapter 2, i.e., access to educational primary goods, fair equality of opportunity, and self-actualization.

To sum up the social justice issues that will be explored in the rest of this chapter: early IQ tests were plagued by a multi-faceted deficiency of objectivity, lacking reliability, construct validity, and ecological validity. All this is in violation of the first principle of institutionalized measurement. Due to the illicit inferences being drawn about the quality of the tests and what they measured, this created a situation in which the tests were misused as part of a tracking system that unjustly distorted the distribution of educational goods, violating all the commitments an educational system ought to have to social justice.

The first large-scale testing infrastructures: scientific racism and the cult of efficiency

> Educational psychologists were, in a sense, victims of their own success. From the mid-1920s on, they became more involved in the continued invention and refinement of tests and less engaged in searching for fundamental new insights into the nature of learning. They took for granted and helped perpetuate plans for school improvement that relied on ever more sophisticated schemes for differentiating between and among individuals. But having found a technology

> that could be applied and tinkered with endlessly, they generally avoided questions about the value and necessity of sorting students in the first place.
>
> -Ellen Lagemann (2000, pp. 93-94).

Psychologist Henry Goddard was the first person responsible for the institutionalization of Binet's tests in America. From the first instance of its use, one could find throughout all the hereditarian theoretical accouterments that came to characterize IQ testing in the US for the next fifty years (Zenderland, 1998). Working at an institution for the "feebleminded" in Vineland, NJ, Goddard, a psychologist in a world run by medical doctors, brought in Binet's test as a means for classifying the various types of patients that ended up in such institutions. This was a problem on which doctors had been unable to reach a consensus, but Goddard's solution of basing the categorization of "feeblemindedness" in terms of scores on Binet's test carried the day (along with his newly invented term, "moron," a label for one of the tricky categories 'just below normal,' with 'idiots' and 'imbeciles' occupying even more 'degenerate' clinical statuses). Importantly, using the test in this way was not far from what Binet intended, in so far as the test was being used to identify those in need of special treatment. It was, in fact, an article by Binet on the nature of "feeblemindedness" that inspired Goddard. But there was one critical difference: Goddard and nearly everyone else working with these populations considered their patients' maladies as entirely determined by hereditary factors (Gould, 1996; Kevles, 1998; Zenderland, 1998). "Feeblemindedness," wrote Goddard (1914, p. 547) "is a condition of mind or brain that is transmitted as regularly and surely as color of hair or eyes."

The test was not just adopted for its usefulness in helping to classify different treatment groups; it was understood to be tapping into something substantive and essential about the nature

of the individual being tested. It was believed that the IQ test could reveal this all-important trait, which was previously hidden from view and assessable only by guesswork and intuition, which is to say, subjectively. In other words, the test was not taken as a pragmatic and helpful shorthand, as Binet saw it, but as a nearly infallible scientific instrument stamping individuals with a permanent label. As Goddard (1920 p. 1) would later tell his audience at Princeton:

> The chief determiner of human conduct is the unitary mental process which we call intelligence.... This process is conditioned by a nervous mechanism and the consequent grade of intelligence or mental level of each individual is determined by the kind of chromosomes that come together with the union of germ cells.... [It] is but little affected by any later influence... As a consequence, any attempt at social adjustment which fails to take into account the determining character of the intelligence and its unalterable grade in each individual is illogical and inefficient.

The test itself began to be compared with newly invented medical instruments that were revolutionizing medical practice by revealing the once-hidden interiors of individuals, such as the x-ray and the thermometer (which was not standardized well enough for significant widespread use until the end of the nineteenth century). “Like the newer diagnostic instruments in medicine, standardized intelligence tests removed the onus of objective diagnosis from the individual practitioner and placed it on a mechanical instrument whose reliability was founded on scientific consensus” (Brown, 1992, p. 93). Psychologists soon came to enjoy a new status alongside doctors as experts in the use of ostensibly uncontroversial objective instruments, a remarkable turn of events given the previously disreputable perception of psychology as an

academic discipline during the middle years of the century. Mental testing would contribute greatly to the rising prestige of professional psychologists, especially their reception by educators as disinterested scientific experts (Brown, 1992; Sokal, 1990). This positioned both them and their instruments as uniquely immune to criticism. As Goddard (1913, p. 9) put it, exemplifying the tone adopted by the early intelligence testers when addressing their critics: those familiar with the tests, the experts, had "become so entirely confident of their [the test's] supreme merit, that the criticisms that arise from time to time only arouse a smile and a feeling akin to that which the physician would have for one who might launch a tirade against the clinical thermometer."

The power of this analogy to medical technology would dominate the earliest use of IQ tests in schools, where they were administered one-on-one in ways akin to the newly instituted medical and dental exams that began to take place in public schools in the 1890s (Brown, 1992). Public health workers in schools paved the way for psychologists, who in turn came equipped with their own objective diagnostic instruments. IQ tests were first used in schools to diagnose thousands of "feebleminded" children (almost entirely poor immigrants), who were thought to be better off in special institutions than in public schools. One of the key arguments for removing them from the general population was the fear that they would spread their "germ plasm," e.g., spread their defective genes through reproduction—the close relationship between the testing and eugenics movements is a topic that will be discussed shortly.

The countless numbers of immigrant children channeled into institutional care (and otherwise pressed out of schools) through IQ testing is a striking case of metrological injustice resulting from a lack of objectivity—or better, *injustice resulting from pretensions of objectivity*. The reason for this is threefold. First, high-stakes decisions were being made using inaccurate,

biased, and inconsistent testing practices. Furthermore, these testing practices were guided by the idea that the tests measured a form of intelligence that is heritable, unchangeable, and constitutes a dominant characteristic of the mind. And finally, the tests were institutionalized under false pretenses about what they measured and how well they measured it, and were thus put to use in inappropriate ways. In sum, the tests lacked reliability, construct validity, and ecological validity.[7]

Importantly, these first uses of IQ testing in schools were already unlike the uses proposed by Binet, who believed the diagnosis of a child as unfit for school was a sign that the child needed help and could attend school again if given the right form of it. The individuals diagnosed in American schools received a very different kind of label, and were essentially removed from school with no prospect of returning or gaining access to any other educational goods. The fact that these individuals were disenfranchised in this way undermines the commitments an educational system ought to have to social justice. Whole groups were denied access to educational primary goods, which negated the role of the educational system in providing for fair equality of opportunity and conditions conducive to individual self-actualization.

Crucially, the institutionalization of this test-based sorting mechanism was predicated on the appearance of scientific legitimacy—that is, the widespread belief, propagated by the leaders of the movement, that IQ tests measured what they claimed to and measured it well. In particular, the language of technological innovation and scientific expertise, a rhetoric adopted by most psychologists, allowed psychological testing to flourish largely unquestioned. There was a critical social discourse disproportionately small compared to the large-scale social impacts of

[7] That IQ tests from this era lacked reliability and construct validity is discussed in more detail the coming sections. It is a well established fact in the literature (see: Block & Dworkin, 1976; Gould, 1996; Sokal, 1990)

testing. It also played directly into the use of testing predominately as a part of efficiency-oriented reforms, the first instance of a theme that will reappear throughout the rest of this work. As summarized by Brown (1992, p. 7), who offers an entire book on the power of the metaphorical language adopted by the early intelligence testers (who molded those psychologically and intellectually fraught metaphors still in use today):

> The technological language with which the psychologists framed their professional agenda served, in fact, largely to remove their enterprise from the domain of politics and thus from the reach of its strongest critics. The bracketing of divisive political issues also strengthened the apparent (and therefore effective) consensus within the discipline. This removal from the larger political domain of questions of intelligence and social worth, of personal labeling and social assignment, was the most powerful effect of the professionalizing process that psychologists furthered by comparing their work to medicine and engineering. It was not until the early 1920s, several years after intelligence tests were mass-marketed, that a chorus of political dissent arose around the issues of democracy, mental testing, and "educational determinism."

In Chapter 5 it is shown that the same pattern characterized the years following the advent of ETS and the legislating of NCLB. Scientific and technological enthusiasm coupled with the rhetorical power of expertise (as well as a sense among those in charge that social justice was being served) created conditions in which critical discourse about the newly created testing infrastructures was slow to dawn, and when it did—it was too little too late. Yet, while

ETS and NCLB involved the large-scale mass administration of tests to millions, the first uses of the IQ test we have discussed so far were a small-scale affair, involving one-on-one administrations. These intimate forms of testing—where the examiner and test-taker sit alone in a room across a table from each other—easily lent themselves to medical analogies. It would take several technological innovations in psychometrics as well as profound institutional upheavals in the public schools to usher in the creation of large-scale testing infrastructures. These events would lead psychologists to begin comparing themselves to engineers and thus to adopting the language of *social engineering*.

Standardized testing: American-made, industrial strength, and efficient

It was the US military that first provided the opportunity and resources psychologists needed to transform Binet's test into one capable of large-scale mass administration. Spearheaded by Robert Yerkes, the project involved a group of psychologists that included, among others: Lewis Terman, Henry Goddard, and Carl Brigham (inventor of the SAT, discussed in Chapter 5). They met in 1917 in Vineland, NJ to begin work on a way to administer intelligence tests on behalf of the Army to the nearly two million new recruits being mobilized for the war effort. The resulting tests became known as Army Alpha and Army Beta (the former for the literate, the latter for the illiterate). These tests were administered to over 1.75 million recruits during World War I (Sokal, 1990).

Two key psychometric innovations enabled this "industrial strength" repurposing of Binet's scale. Both were due mostly to the enterprising work of Terman and his graduate students at Stanford, whose augmented version of Binet's test, the "Stanford-Binet," was the most widely used and respected test of its day (and it would emerge after the war to dominate a

burgeoning multimillion-dollar testing industry). The first innovation was the expansion and refinement of Binet's questions so that the test could be administered to adults as well as children. The second innovation was turning all the questions into multiple-choice items. Together these innovations made for an IQ test that could be administered to rooms full of test-takers of any age and then scored quickly by rooms full of untrained research assistants who could simply overlay standardized stencils on standardized answer sheets (Samelson, 1990).

These innovations transformed standardized testing from something done on a small scale, like a clinical diagnostic, to something done on a large-scale, not dissimilar to the screening, sorting, and calibration of industrial materials in a factory or as part of an engineering project. The language used by psychologists changed accordingly, from medical analogies to engineering ones. "If the Army machine is to work smoothly," remarked Terman, "it is as important to fit the man to the job as to fit the ammunition to the gun" (quoted in Kevles, 1996, p. 81). The same metaphorical language would soon be deployed in the context of educational reform.

The idea that students were best thought of as materials to be worked on and prepared for a proper (e.g., efficient) fit into the economy is at least as old as the factory systems that encouraged that idea. As the influential nineteenth-century educational theorist Lester Frank Ward stated it, "Every child born into the world should be looked upon by society as so much raw material to be manufactured. Its quality to be tested. It is the business of society, as an intelligent economist, to make the best of it" (quoted in Bowles & Gintis, 1976, p. 125). Or as Edward Thorndike (1922, p. 1) remarked in his seminal textbook on educational measurement: "Education is one form of human engineering and will profit by measurements of human nature and achievement as mechanical and electrical engineering have profited by using the foot,

pound, calorie, volt and ampere." Yet never before had there existed a means of objectively and scientifically "testing the quality" of this human capital. The army-testing program would do more than any other initiative to convince psychologists, educational leaders, and the public at large that such measures of human nature were now in-hand and ready to be institutionalized on a scale rivaling that of other newly standardized measures in industry and engineering. Comparisons were made to engineering feats like those that had been used to build the Brooklyn Bridge, the Panama Canal, and the skyscrapers swiftly rising up throughout American cities—complex world-historical engineering feats that captivated the public imagination.

Eugenics and the pretense of objectivity

The army-testing program provided the intelligence testers with an unprecedented amount of data, which was published in a variety of places to much effect. The massive official volume, *Psychological Examining in the United States Army*, was put out by Yerkes (1921) and company. It led to dozens of popular reinterpretations, such as Brigham's (1923) *A Study of American Intelligence*. In all these publications, the results were consistently presented as being in support of ideas associated with the Eugenics movement, then at the height of its popularity worldwide (Kevles, 1996). Essentially a form of scientific racism used to buttress theories of Social Darwinism, the Eugenics movement posited a hierarchy of races, with the "pure" Nordic races at the top, down through a gradation of "lesser" races that bottomed out with Africans. It was a hierarchy of intellectual and moral goodness that was fixed in place by the laws of nature. Understanding the structure and functioning of human genetic endowments would allow for the

improvement of the species through selective breeding and other reflective interventions into the reproduction of human biology.[8]

In the hands of the intelligence testers, this is exactly what the army's massive data set showed to be true. The results were organized along racial and ethnic lines and displayed precisely the distribution that would be predicted by a eugenicist: a gradation of intelligence reflecting the hierarchy of races. However, both the instruments used and the ways in which the data were analyzed and interpreted fundamentally lacked objectivity. Far from being "proven," this data in fact lent no support at all to the hypotheses of eugenicists. But it was the *appearance* of scientific legitimacy that mattered, and the intelligence testers had accomplished that; their ostensibly scientific, large-scale quantitative studies lent an air of rigor and legitimacy to eugenic theories. Of course, *they believed that their instruments were objective and that their theories of intelligence and heritability were true*. This was not a matter of deception or disingenuousness. The problem was that their practices were fundamentally unsuited to addressing the most important aspects of their theory.

The interpretations of the army's testing data exemplify this disconnect between the claims of eugenicists and the tools they used to justify them, as Gould's (1996) enlightening reanalysis of the original data reveals. While the Army had an interest in learning the truth concerning its personnel, the psychologists were prepared to analyze the data in only a certain number of limited ways, which ended up misconstruing the overall results. Item design,

[8] This is not the place to detail the horrendous legacy of the Eugenics movement, described in contemporary terms as intelligence testing's "sister science" (Gould, 1996). The army data was brought before Congress and the Supreme Court as part of the arguments made for legalizing the forced sterilization of thousands diagnosed as feebleminded by IQ tests (a program that would inspire the Nazis). It also informed the Immigration Restriction Act of 1924, which set quotas on immigrants from certain regions based on "proof" that their race was degenerate, effectively sending millions back to Europe as it was falling into the hands of totalitarian governments. This period of widespread institutionalized scientific racism and the role played by intelligence testing has been well documented (Gould, 1996; Kamin, 1974; Nairn, 1976; Zenderland, 1998).

administration procedures, and scoring were remarkably lax, and systemically disadvantaged certain groups, including those who could not read and those with little knowledge of American culture. The data were systematically misinterpreted to downplay the impacts of environment, which could have served as possible explanations for the findings. For example, a wide variety of factors were not controlled for in analyses, including educational background, socioeconomic status, and language of origin. Explaining individual differences in this way — i.e., as due to innate aspects of the individual, and not to the institutions surrounding the individual—was especially problematic because it assumed what it set out to prove, that intelligence is an innate trait that is immune to environmental factors. This kind of argument, which turns a blind eye to issues of economic class, cultural difference, and the inequitable distribution of educational goods, characterized nearly all the research done during the first decades of the IQ test movement (Sokal, 1990; Gould, 1996). Strong hereditarian presuppositions led researchers to ignore or dismiss both possible environmental determinants in IQ test performance as well as explanations for academic achievement that were *not* linked to differences in innate intelligence.

Unfortunately, these forms of argument and analysis eventually found their way into schools. The administration and statistical procedures that accompanied the institutionalization of efficiency-oriented testing practices rendered differences in academic achievement due to socioeconomic class nearly invisible, focusing instead on the quality of the individual child's innate endowment. For example, reporting on one of the first large-scale implementations of testing in public schools, Terman's student Dickson (1922, pp. 33-52), director of research for the Oakland, CA school system, argued that because "mental tests given to 30,000 children prove conclusively that the proportion of failures due chiefly to mental inferiority is near 90 percent," the obvious solution is to track students, "to find the mental ability of the pupil and

place him where he belongs. This policy of segregation is more democratic than former systems because it offers to every child a freer opportunity to use his full capacity." Importantly for this study, in the Oakland slums and poor areas, schools were putting more than fifty percent of students in "limited" or slow classes, while schools in wealthy areas of Oakland put only three percent of students in slow classes and more than fifty percent in accelerated ones (*ibid*.). Yet these results were not taken as an index of the impact of socioeconomic conditions on school achievement, but rather as a sign that the poor were genetically inferior and thus "in their proper places."

The key point here is that standardized testing, eugenicist ideology, and the pretensions of objectivity that came with them, were all brought into American public schools under the banner of *social and economic efficiency*. They were thought to provide what was needed in order to make educational processes as scientifically efficient as their industrial and corporate counterparts (Callahan, 1962). As discussed in Chapter 5, although eugenics would later be abandoned as the scientific backdrop for testing practices, it was nonetheless during this period that large-scale testing infrastructures were first integrated into schemes centering on social and economic efficiency. Here was the birth of efficiency-oriented testing. These practices, and not the overt racism that often played a role in their perpetuation, should be seen as the central legacy of this period. As appalling as the scientific racism surrounding early IQ testing is, to focus solely on this aspect of the origins of testing is to miss most of the lessons from this era that are still relevant to our own. Testing in the name of social efficiency continues, even if its ideological underpinnings have changed. It still relies on practices that often lack reliability and validity; it is still based on theories of education that give primacy to economic efficiency; it is still grounded in an abstract individualism that understands merit and achievement as exclusively

the result of individual endowment and effort. Thus, it is vital to look at the origins of the cult of efficiency, which remains profoundly powerful to this day. Tyack (1979, pp. 180-181) sums up this era during which efficiency-oriented testing was born:

> For leading schoolmen it was mostly an age of confidence inspired by a dream of social efficiency.... Schools [would] serve the needs of the economy for specialized manpower... and the needs of children, which could be scientifically assessed.... Nature-nurture debates might pepper the scientific periodicals...but schoolmen found IQ tests invaluable as a means for channeling children. The [resulting] differentiation of education into tracks represented a profound shift in the conception of the functions of universal education.... Under the spell of the psychologists who dominated educational thought ... educators often failed to see that many problems children faced in school were sociological and economic in character and were, in C. Wright Mill's terms, "public issues" rather than "personal troubles".... It makes little sense to malign the intentions of schoolmen in their campaign to differentiate the structure of schools, to classify students, to socialize in uniform ways. With but few exceptions their motives were good, their belief in the objectivity of their "scientific" procedures manifest, their achievements in the face of massive challenges impressive. But the unforeseen consequences of administrative progressivism become most clear when one looks at the educational experiences of those citizens at the bottom of the social structure....

Justice is efficiency: efficiency is justice

A movement known as "Taylorism" or "scientific management," the brainchild of Frederick Taylor (1912) and his enthusiastic American supporters, revolutionized the way American businesses worked during the first decades of the twentieth century. Taylor's program—symbolized by a stopwatch, and often described as "management through measurement"—was estimated to have saved the new transcontinental railroads billions of dollars and to have more than doubled the output of some factories, while at the same time decreasing their cost of production (Callahan, 1962; Nobel 1977). These remarkable achievements led to a nationwide fervor for efficiency that would begin to equate it with justice and tie it into an emerging national ethos. Theodore Roosevelt (also a supporter of the Eugenics movement) summarized this sentiment in a set of speeches that received national attention:

> Scientific management is the application of the conservation principle to production.... In the factories where it is in force it guards [our natural resources], our raw materials, from loss and misuse. First, by finding the right material—the special wood or steel or fiber—which is cheapest and best for the purpose. Second, by getting the utmost of finished product out of every pound or bale worked up. We couldn't ask more from a patriotic motive, that Scientific Management gives from a selfish one.... You must be efficient. If you are not, you cannot do good to others. You must be efficient. You must never forget for a moment that, so far from being a base theory, it is a vital doctrine, a doctrine vital to good in this country. (quoted in Callahan, 1962, pp. 20, 46)

The Progressive educational reform movement would pick up on this theme and begin explicitly characterizing the minds of students in schools as America's most important natural resource, one in need of conservation and optimization, requiring reforms aimed at improving the efficiency of schools and in turn the overall efficiency of the country (Cremin, 1964). Subsequently, this would lead to the proliferation of educational efficiency experts who brought "scientific management" into the schools (Callahan, 1962; Chapman, 1988; Cremin, 1988; Tyack, 1974). Education efficiency experts began aligning themselves with the same kind of "scientific" practices as industrial efficiency experts, namely knowledge of how the "raw materials" being processed functioned and a way of measuring these materials to assure their proper use. Eugenics and standardized testing appeared to provide just these ideas and techniques: a theory of the limits and affordances of human nature and a way to measure it.

The goals of the efficiency experts were to redesign public education in the image of the "scientifically managed" corporation. Using newly developed "efficiency indexes," their goal was to improve school organization, management, and teacher practice. The experts also had newly developed means for "efficiently sorting students" that would align the outputs of schools with the increasingly complex divisions of labor that characterized the emerging industrial economy (thus improving overall societal efficiency). Both of these practices involved the large-scale deployment of standardized testing, and the widespread administration of tests beginning in the 1920s constituted the first major testing-intensive educational reforms in history. These decades mark the emergence of the education commodity proposition, as well as the first efficiency-oriented testing infrastructures.

In this context the psychologists involved with intelligence testing began to lean heavily on a language emphasizing social engineering, talking about testing as a necessary infrastructure enabling the design of a more scientific and efficient society, which in their minds was therefore more just and democratic.

> "Prediction and control," "human engineering" and "social efficiency" were the catchphrases for American psychology [during the first third of the twentieth century]... and it seemed that biological determinants and hereditarian explanations of human differences were very compatible with the vertical division of labor necessary for an industrialized society. More specifically, the use of mental tests provided an efficient means of classifying individuals in terms of their potential contributions to the social order of the corporate state. Within public education, mental tests were welcomed as an expedient tool for classifying a burgeoning population of schoolchildren, swelled by large numbers of recent immigrants. To bring order out of the chaos, the leaders of the educational establishment, by 1918, had clearly opted for a differentiated curriculum. Vocational education was emphasized at the secondary level—mass education did not mean mass academic education. Thus, there was agreement among psychologists and educators that mental tests had great potential in the schools for appropriately sorting students. Only the students toward the upper end of IQ distribution would be sorted in to academic tracks. (Minton, 1990)

During this period American public schools were expanding at unprecedented rates, especially in growing urban centers. The immense scope of these changes can be seen in the fact that, "between the years 1890 and 1918, there was, on average, more than one new high school built for every day of the year" (Tyack, 1974, p. 183). Millions of immigrants from Europe flooded into major cities at the same time the factory system was causing Americans to migrate from the country to the city in search of jobs. These demographic trends were combined with new (and newly enforced) mandatory attendance laws, and the sheer numbers of students flooding into the schools had educational leaders scrambling for solutions —which the efficiency experts claimed to have.

The basic solution was the same throughout the country, to use standardized "intelligence" testing to sort students into "natural" groupings, thereby allowing administrators and teachers to direct resources more effectively. The dominance of this approach is revealed in reports from the US Bureau of Education on the use of standardized tests in urban schools during the 1920s. It reported that the vast majority of schools used tests in very specific ways and for a limited range of reasons, chiefly for sorting/tracking, identifying "morons," and as a part of vocational guidance counseling (Tyack, 1976, p. 208). Efficiency-oriented testing in this era was such that each school was a testing-enabled "sorting machine" that created a hierarchy of students, resources, and opportunities, a hierarchy that that mirrored that of the broader social system, which in turn was thought to reflect the natural hierarchy of human genetic endowments.

Chapman's (1988) illuminating analysis of testing trends during this era reveals the same thing. Schools were redesigned under immense political and demographic pressures. They were changed over from so-called "Common Schools," where, for better or for worse, every student received roughly the same curriculum, into complex differentiated institutions where different

students received markedly different treatment (from challenging college prep to menial training for factory work). In the vast majority of schools, neither students themselves nor their parents or teachers decided their educational trajectory. Differences in access to educational goods were justified by the "objective" technology of testing and the "scientific" ideology of eugenics. Fates were determined by tests designed by psychologists who never set foot in the schools where they were used. There were, in fact, hundreds of tests on the market, and no accepted national standards of reliability and validity; the test *had only to look like* an IQ test. These tests focused on a narrow range of skills, yet they were thought to objectively reveal the most important innate ability residing in the child. "And to those children, the debates over the validity of IQ measures meant less than the way the tests were used in their everyday lives.... Whether the test had any validity as a test of 'innate metal ability' or not, it would surely have consequences for the pupil—and those consequences could feedback to the child in such a way as to fulfill the prophecy made by the test" (Tyack, 1976, p. 206). Again, here is a clear example of the interface between social justice and objectivity. Pretensions of objectivity, masking a true lack of reliability and validity, resulted in tests being institutionalized *as if* they were measuring a profoundly important property of student's minds. Erroneous and misleading labels were attached to students that radically impacted both their experiences in schools and their futures.

Once labeled with an IQ score, a student would be placed in a classroom with other students who received similar scores and provided curricula and instruction thought to be appropriate for their "level of intelligence." Wrongly labeled students had little chance to prove otherwise (e.g., it is difficult for a student to demonstrate an undetected aptitude for, say, mathematics, if they are never exposed to anything but the simplest math). Moreover, it did not help that the majority of those on non-academic tracks were impoverished minorities—

representing racial and ethnic groups "known" to be inferior (Chapman, 1988). Sorting students in this way was believed to be a more effective way to run classrooms and schools because it freed the teachers' attention by only having to focus on ostensibly homogenous groupings; it was better, as one teacher put it, "because all the 'stupids' are in one class" (quoted in Tyack, 1976, p. 210).

In the language of social engineering, it was understood as more efficient to sort students because doing so separated and classified profoundly different "raw materials," each requiring different techniques and relations of production. Efficiency-oriented testing provided institutional mechanisms that operationalized a eugenically-tinged language of individual differences, which was used to argue that students learn better (and thus that schools run more efficiently) when they are differentiated according to their "natural abilities." Importantly, because the students were classified according an (erroneous) understanding that IQ tests captured their innate natural abilities, the tracking systems were unequivocally ridged; according to principle, certain students were seen as being *legitimately* prohibited from access to educational primary goods, fair equality of opportunity, and conditions conducive to self-actualization (Chapman, 1988).

To clarify: the point here is not that capability grouping are always wrong (they are not), or even that the racist inspiration behind many of these systems was unconscionable (it was). Rather, the point is twofold. First, a tracking system is only as legitimate as the techniques used to determine student capabilities. In this case, IQ tests were taken to be measuring something that there were, in fact, incapable of measuring. Second, any tracking system that undermines the commitments that an educational system ought to have to social justice can be condemned outright, irrespective of the quality of the measures used. In this case, the lowest performing

students were removed from environments conducive to their self-actualization, not accorded fair equality opportunity, and in some cases, were not even provided with the most basic of educational primary goods.

Arguments about innate capability (based on false assumptions about the reliability and validity of IQ tests) were also tied into broader arguments about the social benefits and necessities of stratifying society according to hereditary endowments (measurable using IQ tests). The school hierarchy was seen to reflect the hierarchal division of labor in industrial society, so school systems that scientifically sort students more efficiently route them into careers fitting with their abilities. For philosophers, it is hard to read these arguments, which were ubiquitous in the discourse of educational leaders at the time (see: Brown 1992; Bowles & Gintis, 1976; Sokal, 1990), without thinking of Plato and "the myth of the metals," which occupies Book III of the *Republic*.

This myth was, of course, a "noble lie" suggested by Socrates as necessary to maintain the stability and justice of a society structured hierarchically into three classes: the rulers (whose blood was mingled with gold); their auxiliaries, known as the guardians (whose blood was mingled with silver); and the craftsmen and merchants (whose blood was mingled with brass and iron). The relevance of this ancient theme was certainly not lost on these early modern school reformers, yet there were important differences (Gould, 1996). Firstly, the early-twentieth century reformers did not see their social philosophy as myth or metaphor: they understood it as a scientific truth, supported by both the data produced through intelligence testing and the broad evolutionary theorizing of eugenics. Moreover, while Plato had no pretensions that the ideal republic would in any way be a democracy, the modern reformers did (for the most part) work to characterize their view as one that would lead to a more democratic society. It is instructive to

see how beliefs in hereditarian views of intelligence and in the objectivity of testing instruments were made to cohere with a broadly democratic ethos; it often requires stretching the definition of democracy almost beyond recognition:

> The people who are doing the drudgery are, as a rule, in their proper places. We must learn that there are great groups of men, laborers, who are but little above the child [as measured on IQ tests].... that workman may have a 10-year intelligence while you have a 20. *To demand for him such a home as you enjoy* is as *absurd* as it would be to insist that every laborer should receive a graduate fellowship. *How can there be such a thing as social equality with this wide range of mental capacity?* Democracy means that the people rule by selecting the wisest, most intelligent and most human to tell them what to do to be happy. Thus democracy is a method for arriving at a truly benevolent aristocracy.
> (Goddard, 1919, p. 237-246, emphasis added)

Similar arguments were to be found in the mouths of many other educational leaders, who understood testing itself as a technology enabling a more democratic form of schooling. It is important to recognize that while in retrospect these new testing infrastructures are best classified as efficiency-oriented, at the time they were implemented as justice-oriented reforms. The conflation of justice with efficiency—the idea that a more efficient society is necessarily more just—resulted in structures for optimizing social efficiency, not for administering justice to individuals. So even if the issue of objectivity is put to one side, the guiding vision was one that subordinated individuals to a contrived vision of an ideal social world. While it is true that some

forms of tracking can be understood as beneficial to individuals (because they can thereby be placed in a better position to learn), the forms of tracking at issue here were designed to put and keep individuals in the place most suited for them given their innate endowment. It was, to paraphrase Gould, *testing as a practice for setting limits*, as opposed to *testing as a practice for enhancing potential through proper education.* A full discussion of the issues involved with test-based ranking and sorting is reserved for Chapter 5, where the nature of test-based meritocracies is discussed. Such issues did not fully surface during this period because the ubiquitous lack of objectivity was enough to make even the most well intentioned system for test-based tracking unjust.

Meditations on the birth of testing

> If the misery of the poor be caused not by the laws of nature, but by our institutions, great is our sin.
>
> -Charles Darwin (1839)

The task of this final section is to work toward establishing a broad reflective equilibrium between the historical material just presented and the theory of just educational measurement that is the focus of this work. The theory is composed of the six principles derived in Chapters 1 and 2, as well as the education commodity proposition and the dynamics of efficiency-oriented testing discussed in Chapter 3. Some of these conceptual distinctions and models have already

been referred to in passing above, but this concluding section will more explicitly use the theory to draw out the central social justice issues surrounding the birth of testing.

The first set of issues concerns the justice of pursing objectivity as well as the *injustice of unexamined pretensions of objectivity*. The most basic way in which to articulate the injustice of early testing infrastructures is to say that they violated the first principle of just institutionalized measurement. Early intelligence testers put in place a measurement infrastructure that lacked objectivity. As discussed in Chapter 1, with the lessons of the unjust miller, in some cases a non-objective measurement infrastructure is *worse than no formal measurement infrastructure at all*. In the case of early testing, nonobjective measurement led to the *wrongful distribution* of educational goods and contributed to an educational system that *undermined equality of opportunity* and demanded *the forfeiture of self-actualization* for countless students.

Recall that the principles of just institutionalized measurement begin with each individual's right to objective measurement. The ethical imperative for establishing objectivity in the basic metrics that constitute social interactions is an ancient and perennial theme, seen both in the Bible and in the discourse surrounding the emergence of the metric system in revolutionary France. This ethical imperative was also present at the birth of large-scale standardized testing, as Binet's orientation toward his invention shows.

As discussed in Chapter 3, tests are more complex than meter sticks and scales. Testing always establishes an interpersonal relationship requiring a shared understanding of a wide range of values and norms. The relationships established through testing can be liberating—as Binet thought they would be if they were designed solely to aid the least well off. Binet had *the justice of objectivity* in mind when he invented what would become one of the dominant measurement technologies of the twentieth century. Importantly, for Binet in France with a homogenous

population of children, pretensions of objectivity were warranted—although he was cautious in even these limited claims. But when his tests were repurposed for industrial scale use as multiple-choice tests, what semblance of objectivity existed vanished and possibilities for their use as instruments of justice diminished.

This is one of the main lessons of Gould's incisive critique of the early intelligence testing movement, *The Mismeasure of Man* (1996). Gould demonstrates again and again the sheer lack of objectivity that characterized the early use of mass-administered IQ tests. Yet IQ tests were institutionalized, commercialized, and popularly understood *as if* they were objective. Test results set the terms of the mutual understandings that structured social relationships within the education system. Testing functioned with a kind of epistemic authority, justifying the treatment of individuals according to the categories of the larger bureaucratic structure—enabling "scientific management," also known as "management through measurement."

As discussed above, these early measures were not reliable: they did not function in an accurate, consistent, and unbiased fashion. Nor were they valid: they did not differentiate individuals according their heritable, innate intellectual endowment. The tests failed to do this not only because this kind of "intelligence" does not exist (Gardner, 2011; Gould, 1996). Even if a single unified intelligence *did* exist, the tools and techniques of the early intelligence testers were too inaccurate and unscientific to get at it. But more important than the "bad science" that resulted were the ethical consequences of the (unacknowledged) inadequacies that plagued the early IQ movement's measurement practices. Undermining the first principle of just institutionalized measurement put in motion a cascade of injustices throughout the systems that rely on whatever "false measures" are instituted.

A testing infrastructure lacking objectivity but institutionalized as if it were objective will almost always undermine social justice in schools, especially if it is tied to high-stakes consequences, such as students being sorted into capability groupings. False assertions of objectivity create shared misconceptions about reality. The relationships engendered through testing are ones in which there is an asymmetry of power, both epistemic and bureaucratic. In the case at hand, pseudo-objective tests created shared misconceptions, distorting the reality of what individuals were thought to be capable of, and more profoundly (and inappropriately), of what individuals were "worth." It is hard to estimate the impact of building an identity (as a student) and a professional life (as a teacher and administrator) around a system of categories and practices that are not what they claim to be. Moreover, there is a tendency for tests to reinforce their institutional function. This was an idea discussed in Chapter 3, where it was shown that institutionalized testing infrastructures tend to create school cultures that will perpetuate the status quo test and its categories. Thus the likelihood a shared (pseudo)reality will be disrupted or countervailed decreases as the institutional importance of the test increases. These are lessons about testing that are still relevant, as Chapter 5 will show in a discussion of the conditions that undermine objectivity in federally mandated accountability testing (e.g., erroneous test design and scoring; cheating; test prep).

The ramifications of not meeting the first principle of just institutionalized measurement are so serious that the rest of the framework appears almost unneeded. Like Rawls's principles of justice, if the first one is not met (e.g., if there is not a system of equal liberties for all) then the rest are not worth discussing (e.g., what good are considerations of equal opportunity or distributive justice if basic freedoms are not secured). Indeed, when considering the birth of modern testing, concerns about the relevance and benefits of measurement, equality of

opportunity, and self-actualization seem almost entirely out of place due to the egregious lack of objectivity and the manifest racism of the guiding social philosophy.

It is sometimes said that even in these early decades testing did bring benefit to some, equalize opportunities in certain contexts, and catalyze self-actualization for a small number—these cases mostly refer to individuals who were "lifted out" of social contexts because of gifts detected through testing. These cases are important. But the fact is that if the test is not objective, then it provides a false justification for the resulting distribution of benefits. Benefits are actually being arbitrarily distributed (when reliability is poor), or distributed according to some non-disclosed or unknown criteria. That some individuals benefit from a non-objective system is but one manifestation of the injustice of non-objective measurement; there is no good reason for benefits to accrue to some instead of others—it is either the roll of the dice, or more often, a system that covertly and systematically favors some over others. It bears repeating: if the first institutionalized standardized-testing infrastructures teach a moral lesson, it is centered on *injustices accompanying pretensions of objectivity*.

However, there are also lessons to learn from this era about the first manifestations of the education commodity proposition and the earliest forms of efficiency-oriented testing, which were not fully formalized due to a lack of objectivity in testing practices. Yet despite this lack of objectivity in early IQ testing efforts, educational reformers, policy makers, and school leaders used tests to simplify representations of student ability and decision protocols concerning the value of educational processes. Most starkly, by claiming to be measuring an innate biological endowment those administering these first IQ tests displayed a fundamental confusion about the epistemological status of psychological measurement instruments, which leads to the education commodity proposition, and in turn the hegemony of efficiency-oriented testing. Hereditarian

views of intelligence are, in a way, the most primitive instance of the educational commodity proposition. The test score is used to quantify the "worth" of the individual; like testing the tensile strength of steel or the caret of a diamond (two analogies used by Yerkes). Just determinations of the strength of steel are used to inform its price and how it can be put to use. Likewise, the earliest testing infrastructures were used to determine the value of the student as an economic investment relative to assumptions about their future functional fit in the social system. These earliest ritualized practices instituted by the cult of efficiency foreshadow all the injustices that testing would come to perpetuate well into the twenty-first century.

Efficiency-oriented testing began as an idealized vision of an educational system in which psychological technologies could be used to orchestrate a harmony between social efficiency and biological determinism; it was self-consciously utopian. This vision of "justice as efficiency" would guide the widespread institutionalization of a very specific approach to testing—using multiple-choice tests to sort students and to monitor and facilitate the social efficiency of the schools. By the 1930s, this ideal was being forged into a reality through the use of testing practices that were overtly discriminatory and lacked even the most basic requirements for objectivity. The weaknesses of their tools and the errors of their ideology did not dampen the dream of testing-enabled social efficiency that had captivated the leadership of the early IQ testing movement. However, it would not be until after the Second World War that testing technologies would begin to make good on their dream of a national testing infrastructure for social efficiency.

5: Social justice and the rise of national testing infrastructures

> Popularization and multitudinousness, then, were not less characteristic of American education during the twentieth century than they had been during the nineteenth. But they were now joined by a third distinguishing feature that lent a rather different character to the enterprise, namely, *politicization*, meaning, as in the broad Aristotelian sense, the increasingly direct harnessing of education to social ends.... In part, it was didacticism [and related forms of testing] that underlay the effort... there was a tendency throughout the Western world during the nineteenth and twentieth centuries to shift from paradigms of study to paradigms of instruction. For a wide variety of groups, education in a forcefully didactic mode became, not merely a complement to politics, but a form of politics and increasingly a substitute for politics.
>
> -Lawrence Cremin (1988, pp. 651-652).

> We shall one day learn to supersede politics by education.
>
> -Ralph Waldo Emerson (1837)

This chapter addresses the social justice issues that characterize contemporary testing infrastructures. Continuing with the task of drawing lessons from the history of testing, the first section below deals briefly with the history of the Educational Testing Service (ETS). Established in light of a vision of justice-oriented testing, ETS was undertaken in the name of

objectivity and in the pursuit of a science of education. ETS would build the first national testing infrastructure, touching the lives of nearly every college bound student. The size and impact of "the big test" would shape the character of educational institutions and change the face of the testing industry (Lemann, 1999). A true science of education would remain elusive (Lagemann, 2000), but ETS would pull off the nationwide institutionalization of a test-based meritocracy. On the one hand, this was a social justice boon. Not only did ETS correct the total lack of objectivity that plagued the early IQ-testing movement, it replaced an aristocratic educational system based on privilege and family lineage with an ostensibly meritocratic one. On the other hand, it was the first efficiency-oriented testing infrastructure of national scope; ETS had built a meritocracy that was designed to channel resources and opportunities toward those who performed well on a very specific and narrow set of tests. These tests were originally designed mainly to identify and promote individuals with skills thought to be of great "national value"—specifically, scientists and engineers in the context of the Cold War.

Yet this kind of test-based meritocracy is prone to injustice. It leads to inequalities between individuals that are much too great, disregarding Rawls's difference principle (and any similar principles concerning distributive justice). These inequalities are justified in terms of high-stakes testing practices that, while far more reliable than the early IQ-tests, lack robust construct validity and end up create systematically distorted forms of self-understanding on the part of participants. This undermines an essential condition of free and equal citizenship. ETS created a profoundly new kind of educational system in which opportunities and rewards could be allocated "scientifically," on the basis of test-based determinations of merit, and in light of national personnel requirements. Because of its success, ETS now exists largely unchallenged as the biggest testing company in the world. The main focus of this chapter, however, takes off

from where ETS's remarkable (if double-edged) achievements have left the public education system: captivated by the idea of large-scale objective testing backed by the prestige of scientific and governmental authority.

During the first decades of the twenty-first century, the US federal government undertook the largest testing-intensive educational reforms in history. Beginning with the reauthorization of the Elementary and Secondary Education Act, more commonly known as No Child Left Behind (NCLB), they have continued through Obama's Race to the Top (RTT), and into the still-emerging movement around the Common Core Standards and Assessments (CCS&A). These are all examples or aspects of large-scale efficiency-oriented testing programs. NCLB, the only reform effort that has been fully implemented, is generally understood to have failed, leaving in its wake a multitude of injustices, from dishonesty and cheating to faulty test design and scoring (RAND, 2011). But even in cases where objectivity was secure, the testing infrastructure radically limited curricular options by incentivizing "teaching to the test," leaving students with educational experiences that were at best irrelevant, and at worst harmful and antithetical to their self-actualization.

NCLB serves as a case study in how efficiency-oriented testing can lead to: 1) a *decline in objectivity* (as high-stakes pressure leads to cheating, data manipulation, and lessening standards in test design); and 2) the *inefficiencies of injustice* (as lessening objectivity combines with the mounting costs of surveillance and enforcement, leading to a decline in overall quality and an increase in inefficiencies in the handling of money and time). Despite these obvious drawbacks, the new CCS&A do not appear to be a significant departure from the approach to efficiency-oriented testing that began with NCLB (Ravitch, 2013).

It appears that schooling in the near future will continue to be profoundly shaped by efficiency-oriented testing practices. The entire infrastructure—from test-taking to scoring and analysis—is going online, leading to a mass expansion and proliferation of data processing and storage systems (Colins & Halverson, 2009). Yet just as technological innovations are poised to change testing infrastructures in profound ways, trends in government and industry are poised to transform the educational system itself into a marketplace. As marketization and privatization combine to refashion the way educational goods are provided, testing will likely play an even greater role in attempts to foster standards and exercise quality-control in educational marketplaces. This comes with its own ethical problems, especially because students are caught up in the market as *both* products and consumers—a vulnerable position that makes them doubly prone to being objectified as a part of test-mediated educational relationships.

On the ethics of national testing infrastructures

Just as in the previous chapter, it is necessary to first present an overview of the central ethical issues that will occupy the coming sections. The issues to be discussed here stem from the ones presented in Chapter 4, which were built off the theory of just educational measurement developed in the first three chapters. It was shown that the early the IQ testing movement had the following issues: 1) testing practices characterized by *a multifold lack of objectivity,* including a lack of even basic forms of reliability and construct validity, and; 2) *institutional misuses of testing*, including their use for high-stakes decisions despite the problems noted in 1, and their use as part of an unjust tracking system (based in part on ideas associated with the problems in 1 [e.g., innate intelligence]). The same set of issues are largely at play in ETS and NCLB—

concerns about objectivity and institutional misuse—but with a difference of emphasis, severity, and ethical impact.

ETS solved the problems with reliability that so radically undermined the early IQ-testing movement, putting in place an infrastructure that would secure reliability through the "mechanical objectivity" of automated test scoring. ETS's approach to testing also lessened the intensity of the problems related to construct-validity by stepping back from claims about both innate intelligence and heritability. The SAT was explicitly not intended to measure innate intelligence, but instead was proposed as a measure of *aptitude*, and more remotely, *academic merit*—a combination of intellectual endowment and effort in school. While the claim to be measuring scholastic aptitude was less scientifically outlandish and ethically insidious than the claim to be measuring heritable innate intelligence, it was still incorrect. This meant that for years college placement justifications based on the SAT were problematic because there simply is no single psychological structure, trait, or property of scholastic aptitude. Scholastic aptitude is not what the SAT measures — a fact that has become increasingly clear thanks to advancements in psychometrics (Lemann, 1999).

Today, the notion that the SAT measures scholastic aptitude has been abandoned (in theory, if not yet fully in practice) and the test is merely taken as an a-theoretical index of the "likelihood of success in college." In other words, the test is only deemed valuable in terms of its predictive validity. This reliance on predictive validity means the test has no referent *per se*; the SAT does not measure anything except how students do on the SAT, which is then shown to have modest simple correlations to first year college grades. As discussed further below, what this means is that the SAT is not *about* anything; it is certainly not about a domain of competencies as broad as "college readiness" or "scholastic aptitude." If we are honest about

what the SAT measures (i.e., how skilled one is at taking the SAT) then the justifications for its use lack force and sound strange: "You are not being admitted because you didn't do well on a test that we are told correlates with success in college, although we are not sure why these correlations hold or if we value the skills measured by the test."

While the correlations between SAT scores and first year college grades are better explained by factors such as the presence of a family that can afford SAT test-prep (*Ibid*), the main the problem here is not the complexities and debates surrounding the predictive validity of the SAT. The issue is a deeper epistemological one: predictions of success or failure in college based on SAT scores cannot be *explained* in terms of knowledge yielded by the test. They are predictions culled from statistics, not predictions justified by explanations making use of the psychological constructs objectively measured by the test. This is a complicated point, and one that will be elaborated further below. Now, however, it is important to point out the broader implications of what this means: despite important advances that greatly improved the test's reliability, the SAT is still problematic in terms of objectivity, insofar as an objective practice must be *about* something. In Chapter 1 it was argued that in science in general, and in measurement more specifically, an objective practice is one carried out reflectively and explicitly in reference to something (usually the point of conducting measurements is to improve our understanding of that something, but this is a separate issue, i.e., that tests like the SAT do not advance basic research in psychology).

As was the case with IQ testing, problems with objectivity lead to problems with institutional misuse. This is because there are important relations between what a test is claimed to measure and how it is institutionalized. The main problem with the SAT, as will be discussed in the sections to come, is that high-stakes decisions require more than good reliability. Make no

mistake: reliability is critical because its absence disqualifies any and all tests from high-stakes use. The SAT is admirable on this count, truly raising the bar in terms of reliability. As such, it does away with a whole range of social justice problems, such as those that arise when individuals' educational futures are changed as a result of *misclassification* (e.g., rampant errors in scoring and statistical analysis; highly variable conditions of administration; overt cultural biases in item design; etc.). But high-stakes decisions require more than mere reliability. Justice requires that a high-stakes test fulfill the fully robust definition of objectivity offered at the end of Chapter 1—involving a combination of reliability with construct and ecological validity. Construct validity requires some knowledge of what the test is about—knowledge about *what* it measures. This informs the kinds of claims that can be made based on test results and thus determines how the test ought to be institutionalized. Ecological validity raises questions about whether what the test measures is appropriate given its institutionalized use. The point here is that what a test is about—what it measures—plays an essential role in the justification of institutionalized test-based decisions, especially high-stakes ones.

Take the classic example of a measurement-based decision at an amusement park: "You can't go on this ride because this measuring stick (which we both know objectively represents your height) says you are not tall enough." In this case, the park employee justifies their decision to exclude you on the grounds of having knowledge of your height and knowing in turn that the ride's safety mechanisms will not sufficiently protect you. Or take another high-stakes testing situation where the test is directly about the quality that matters: "You can't fly this plane, because you failed the flight simulator test, which is about your ability to fly this plane." Or a less extreme case that occurs in some form everyday in schools: "You can't be in this reading group, because you were unable to read the books assigned in the lower skill group." Each of

these instances of measurement-based inclusion/exclusion decisions involve inferences implicating knowledge acquired through measurement about qualities that matter for the situation at hand—riding a ride, flying a plane, or advancing to a more challenging reading group.

There is, of course, good predicative validity for these tests—if you are too short, you will fall out; if you can't fly you will crash. But these predictions can be *explained*; they are not merely correlations. We know how what is measured interfaces with the situation at hand. One doesn't need to put kids on the ride or novices in the cockpit to get statistics on predictive validity for these tests. The measure and our knowledge of things are good enough. Of course, psychological measurement is harder, and predictive validity is always less robust in open-systems and when non-physical measurement instruments are involved. Furthermore, predictions are problematic in any science, but especially in the human sciences, which deal in actions, motives, and reasons, not just simple physical causality (if there even is such a thing). This is why explanations are so important in the human sciences, and why predictions can hardly ever be expected to pan out (Bhaskar, 1998).

The example of the reading group assignments is instructive here. Reading a certain set of books is a direct measure of ones ability to read that set of books. While not a test in the traditional sense, ongoing tasks such as reading and discussing a handful of books are typically going to provide a much better sense of a student's reading ability than any one-off reading comprehension test. If I have discussed a handful of books with you under the right conditions, the decision I make about your reading group placement is not best thought of as a *prediction;* it is a recommendation based on my direct knowledge of your reading ability. Importantly, I can never really know if you will succeed in the higher reading group, no matter how good my

knowledge of your abilities or how perfect my reading comprehension test. There are so many other factors impacting success (e.g., your interests in other subjects, problems at home, illness, etc.) that predictions are never a sure bet, especially in complex human systems like schools. This is why, when it comes to testing, the need to understand and explain is greater than the need to simply predict.

As the aforementioned examples demonstrate, a reliable measure with good construct and ecological validity provides information that allows us to *explain our predictions*—not necessarily to be right about them (nothing can provide that). These explanations are built based on diagnostic information stemming from an articulate understanding of how what is measured relates to the situation at hand. This kind of diagnostic information allows us to justify high-stakes test-based decisions using explanations that make reference to the constructs and properties measured by the test. Conversely, the set of numbers representing a student's SAT score provide almost no information that can be used in justifying decisions about college success, *except in the form of predictions that cannot be explained*. We know the SAT has some predictive validity but cannot adequately explain why we think any given student will or will not do well in college based on an understanding of what the result means about his or her skills and capabilities.

This reliance on predictive validity (and related explanatory inarticulacies stemming from a lack of construct validity) reflects a broader truncation of concerns about objectivity. Recall the fully adequate conception of objectivity outlined in Chapter 1 (which is based on generally accepted psychometric standards: AERA, 1999), wherein for a test to be objective it must be highly reliable and also be *about* something, having robust and explicit construct validity. In terms of this broader definition, much of contemporary testing practices (such as the SAT and

the practices following after NCLB) appear to be guided by a truncated notion of objectivity, which is conceived as merely a combination of reliability and predictive validity. The idea being that there is no need to have a clear psychological construct targeted by a test, instead you only need an instrument that is reliable and evidence that its results correlate with outcomes that matter institutionally. That is, instead of theory and research about *what* the test measures, there is a focus merely on statistical aspects of *how* the test functions (i.e., mainly statistical studies of internal reliability and predictive validity). This a-theoretical and heavily quantitative approach is, in a sense, the opposite of the hyper-theoretical and statistically primitive approach of the early IQ testing movement. Whereas testing industry researchers once made outlandish claims about the psychological properties their tests measured, while presenting very few forms of statistical evidence, now they make almost no claims at all about the nature of the psychological realities they are dealing with, while presenting endless streams of research on the statistical reliability and predictive validity of their instruments. As discussed in Chapter 4, the IQ testing movement led to injustices stemming from *pretensions of objectivity*. This chapter explores how more recent forms of testing have led to injustices stemming from *truncations of objectivity*.

This truncation of objectivity goes beyond simple claims that the SAT has limits and doesn't tell us everything. This is, of course, why no college admissions official would ever say they base their decisions entirely on a test score. Claims like these are true when it comes to those who get accepted; their transcripts are looked at from various angles. But it is a bit of a disingenuous position overall, as almost all college admissions offices will admit to using SAT cut-off scores at some point during the review process, thereby excluding a large number of applicants simply on the basis of their test performance alone. But no matter what the college admission practices are, any high-stakes decision using a test that does not fulfill the full

requirements for objectivity will result in unjust distortions of educational practice. In this case, as discussed in the sections below, the result was an unconscionably narrow test-based meritocracy, which distorted the perception of educational value and thus drew the focus of individual efforts away from broad educational goals and toward a shallow focus on test-prep and institutional advancement.

Moreover, truncating objectivity means that what a test is *about* will become increasingly unimportant as a factor in testing practices. This often occurs, unfortunately, as a test's institutional uses become more important. Truncating objectivity tends to deflate broader concerns about the relations between construct and ecological validity (i.e., about possible institutionalized misuse). Perhaps it is simply easier to relegate questions about the value of the test to the technical realm of reliability and predictive validity statistics than to open up more complex evaluative and ethical discussions about what the test measures, and what it ought to measure given how it is institutionalized. Eventually this kind of systematically distorted communication about the meaning and value of testing leads to declines in the basic reliability of testing practices, especially in the form of cheating and intensive test-prep. This is discussed below in terms of the *decline in objectivity* and related *inefficiencies of injustice* that accompany much of high-stakes testing. Again, high-stakes testing atmospheres will often distort representations of educational value, as both the money and time invested in intensive test prep, for example, are by any reasonable standards terrible investments to make in low-quality educational goods. This leads to the second set of historical examples discussed in the chapter, the recent history of test-based accountability.

As will be explored in the second half of this chapter, NCLB took a step backward from ETS in terms of issues discussed under 1 above. Reliability differed rather drastically between

states (which is a serious problem if between state comparisons are being made), and was very poor in some school systems (usually those that could not afford to work with competent test providers). But more important than these simple failures in test design and implementation, was the widespread and broadly supported test-prep curricula, often so intensive that it invalidated the testing practices themselves (Koretz, 2008). The *decline of objectivity* that results from intensive levels of test-prep curricula is important to understand (and also applies to the SAT): the vast majority of tests are not intended to be about how well students have been taught to take the test, they are intended to be about how well a student knows certain domains of knowledge or can exercise certain capabilities, *in general and on the whole*. Testing is meant to *sample* the quality of student capabilities, not become the sole focus of student learning. Tests lose their intended meaning when they become the sole focus of instruction; the claims we want to make using the results they yield are no longer valid. Moreover, with NCLB, tests that were designed and intended to measure student capabilities (and could barely do that well enough) came to be used as if they are were measures of teacher and school quality. This is all discussed below in terms of the *decline of objectivity*.

The impacts of this decline raise the issues discussed under 2 above: NCLB created institutional practices that began to radically disfigure student experiences and limit learning outcomes. This was perpetuated and accelerated by the misrepresentation of educational value resulting from polices that institutionalized testing as the be-all-and-end-all of schooling. One way to think about this nearly obsessive focus on testing is in terms of the truncation of objectivity just discussed, i.e., when a test is reduced to reliability and predictive validity—when it is no longer strictly *about* anything. When a test has no meaningful referent, the test itself must become the focus of attention. Because the test does not measure something, or rather because it

measures only performance on itself, there is no way to get better on the test aside from focusing on how to take the test. In some cases, this focus led to mindless test-prep that actually decreased access to basic educational goods. In others it resulted in the forfeiture of student self-actualization due to unreasonable limitations of curriculum. Under the NCLB testing regime, the overall value of educational goods declined and the system was locked into a downward spiral—the key stakeholders who "believed" in efficiency-oriented testing as a panacea prescribed more of what made the system sick in the first place. This is all discussed below in terms of the *inefficiencies of injustice*.

To summarize the ethical issues that will occupy the rest of this chapter: during this era testing grew to have enormous importance in American education and there were some gains in the name of justice-oriented testing. But problems with objectivity and institutional misuse remained, even as testing infrastructures were first being built on a national scale. These infrastructures were mainly designed in terms of a truncated conception of objectivity—tests were made reliable, and sometimes predictive, but what they were actually *about* became increasingly less important, as did questions about whether what they were about was appropriate given their institutionalized use. The resulting testing practices led initially to a narrow and distorted test-based meritocracy and then to a counterproductive and inefficient test-based accountability system. In both cases, excessive significance was placed on tests that were not truly objective, that were often full of irrelevant content of questionable educational value, and that were of no benefit to the student, either in terms of diagnostic information or in advancing their acquisition of basic educational goods. Consequently, these forms of testing have distorted perceptions of educational value, as well as impacted the distribution of educational

goods, equality of opportunity, and possibilities for self-actualization, which are the minimal commitments of a just educational system.

ETS and the first national testing infrastructure: from an aristocracy to a meritocracy

> In January, 1934, James B. Conant [president of Harvard at the time and soon to be a director on the Manhattan Project] instructed Henry Chauncey [who would be the first president of ETS] to use the SAT to select ten young men from the Midwest for scholarships to attend Harvard that fall... After three successful rounds of scholarships [and the participation of the rest of the Ivy League], Conant took the much more dramatic step of proposing publicly that a new national testing agency be created to operate all the leading standardized educational tests.... The agency would become the home of research on and development of future tests.... It would form at least the embryo of a national personnel system, and it could affect exponentially more lives than those of a few thousand aspiring Ivy League students.
>
> -Nicholas Lemann (1999, pp. 39-40)

The SAT or Scholastic Aptitude Test was created by Carl Brigham in 1926. It was literally just the Army Alpha test (discussed in Chapter 4) made more difficult so that it could be used with high school graduates, as opposed to army recruits, children, or the "feebleminded." The SAT was using the same basic technology of testing—mass-administered, multiple-choice—

and riding the wave of the perceived success of the army's testing efforts. Brigham worked with the College Board and other educational organizations to get the SAT into a wide range of schools, from graduate law programs to prep schools (Lemann, 1999; Nobel, 1978).

At the same time, Brigham (who was mentioned in Chapter 4 because of his eugenically based interpretations of the army data) was undergoing a serious intellectual change of heart. Brigham recanted his views before a meeting of eugenicists and disavowed the conclusions of his best-known book. He began explicitly distancing the SAT from the IQ tests that were its origin. There is, for example, no mental age conversion scale on the SAT, just a number representing scholastic aptitude and, importantly, no claim to measuring innate intelligence. Brigham believed that the leaders of the IQ movement had committed a mistake in thinking that by producing a reliable test result they were thus measuring a biological trait in the brain, a belief for which there was (and is) no proof. As Brigham (1934, quoted in Lemann, 1999, p. 34) would put it, intelligence testers committed "one of the most glorious fallacies in the history of science, namely, thinking that the IQ test measures innate intelligence purely and simply without regard to family or schooling. I hope that nobody believes that now. The test scores very definitely are a composite including schooling, family background, familiarity with English and everything else.... The 'native intelligence' hypothesis is dead."

Only two decades before, Brigham was among the most powerful men arguing in favor of the native intelligence hypothesis. But now he began to argue for a very different view of testing. He had come to understood testing as a "method of interview" rather than a "measure of some mysterious power"—an index of certain aptitudes, *not* of innate genetic endowment (*ibid*.). This understanding of testing captivated the imagination of James B. Conant just as he was making plans for a federally supported and philanthropically funded consolidation of the testing

industry. The SAT was to be made the focal point of this new organization. However, Brigham began to express concerns about the potential dangers of a single organization owning the rights to so many tests. Of special concern to him was the idea that such an organization would inevitably become more interested in promoting the use of its tests than in researching their effectiveness or making different and better ones:

> A new organization solely for the dissemination of present knowledge concerning testing and the promotion of testing programs would be difficult to justify. It is the writer's belief that the present testing movement carries the germs of its own destruction. The cure is to be found in research and more research. The provision for an extensive research program will prevent degeneration into a sales and marketing group.... We are today approaching the ultimate state in which the testing movement may take on the aspects of a religious crusade.... It is easy for a powerful organization to set up false ideals. The new organization must be contrived so that it will always remain the servant of education and never become its master. It should inquire into the nature of values but it must not determine those values. (Brigham, 1937, p. 756)

Brigham would die at the age of 52, only a few years after these words were published. While the mission statement of ETS would try to incorporate the warnings of the man who invented its central technology, the history of the organization would constantly prove him prescient in his worries. In a story with remarkable parallels to Binet's, Brigham would have been dismayed at the way his invention was ultimately put to use. Brigham did not offer an ideal

of justice-oriented testing as explicit as Binet's, but he did offer clear explanations of how testing can become an instrument of injustice. He clearly saw the fallacy of treating psychological measurement as if it were physical measurement, the most egregious case of which was the "native intelligence" hypothesis. The ethical significance of this was not lost on Brigham. He saw the power of testing to set the terms from which others choose, to exercise a kind of anonymous power over the shape and character of education. In a letter to Conant just before his death, Brigham warned about the tendency of testing to result in the devaluation of the unmeasured, or as discussed in Chapter 3, the tendency of testing to limit what counts as valuable education and to distort perceptions of educational value:

> If the unhappy day ever comes when teachers point their students toward these newer examinations, and the present [testing] procedures get a grip on education, then we may look for the inevitable distortion of education in terms of tests. And that means that mathematics will continue to be completely departmentalized... that the sciences will be come highly verbalized... that languages will be taught without reference to literary values, that English will be taught for reading alone.... (1939, quoted in Lemann, 1999, p. 40)

The birth of an American meritocracy

Conant heard Brigham's warnings and waited until he died to move forward with the creation of ETS, placing the SAT at the center of the new organization. Perhaps having overseen other major projects that wielded potentially dangerous scientific technologies (e.g., the atomic bomb), Conant trusted in his powers to create a benevolent institution (Conant, 1970). Concerns

about the impact of testing on schools or about the probability of a consolidated testing agency squelching research and development were small compared with the possibility of using ETS to reshape the American educational system as a whole. In a series of increasingly radical articles written for the popular press, Conant used his visibility as an educational leader to call for a Cold War-inspired American *meritocracy* that would be a blend of science, efficiency, technology, and social justice (Conant, 1943). This vision of justice as testing-enabled social efficiency is what Conant saw in ETS, and it involved a virtual leveling of the previous system of higher education in America. Conant's vision made testing the dominant institutional mechanism for transforming an aristocracy of wealth and family influence into a meritocracy of aptitude and effort. He was aware of how radical this vision of educational meritocracy was, but went one step further to suggest that such a system of education could compliment a system of estate taxes that would level all family inheritance—"to confiscate (by constitutional methods) all property once a generation" (1943, p. 43). This would lead to the reinvesting of large private fortunes into a common meritocratic educational system (Lemann, 1999).

Short of this unabashedly utopian vision, Conant believed he had worked out an immediately workable system in which the *personnel needs* of the nation could be determined and then strategically met by the educational system through a well-designed testing-and-scholarship incentives program. This national manpower-sorting machine would funnel the best students into the sciences and other areas necessary for national security while filtering the rest into other economically productive positions. The National Defense Education Act of 1958 was the first of several federal interventions into the public education system in the middle decades of the twentieth century that contributed to this vision (Spring, 1989; Nobel, 1978). It provided for the development of new research-based science and math curricula. It also created a related

testing infrastructure, both to demonstrate the effectiveness of the new curricula and to allot placement in new programs on the basis of merit (i.e., merit as *defined by performance on standardized tests*). As Conant saw it, ETS was intended to take "a deep-seated wish in world history—a governing elite selected on the basis of merit, not parentage—and the most deep-seated wish in American culture, opportunity for everyone, and conjoin them in a new kind of educational system" (quoted in Lemann, 1999, p. 52).

It is important to understand that for executing this vision Conant would have nothing to do with tests of heritable innate intelligence or tests of learned skills or capabilities (*Ibid*). He reviewed many different tests to serve as the flagship for ETS, and was drawn to Brigham's SAT because it claimed to be a test of *aptitude.* On the one hand, after the atrocities of WWII, Conant was aware of the deeply unethical implications building a system such as he envisioned around tests that claimed to measure heritable innate intelligence. On the other hand, tests of learned skills would privilege those from already privileged backgrounds, which was exactly the aristocratic perpetuation of educational access that the new testing infrastructure was supposed to undue. Aptitudes, so Conant believed, are not genetic and deterministic, nor do they merely reflect a person's background and the opportunities they have had. Conant was looking to have ETS design and deploy tests (starting with the SAT) according to a view of the mind where individuals are endowed with various aptitudes, which they can either put effort toward developing or not. Putting in effort to develop aptitudes results in an individual's accrual of *merit*. To adapt an equation from the utopian sci-fi book *The Rise of The Meritocracy* (Young, 1958), where the term *meritocracy* was popularized, the idea was: aptitude + effort = merit.

According to this way of thinking, while a student might not do well on the SAT, this is not a wholesale condemnation of his inherited intelligence nor does it speak to any of the

student's moral qualities (as many believed IQ tests did). Doing poorly on the SAT merely means that the student has poor scholastic aptitude, not that they are genetically dumb. It implies only that they may not be a good candidate for college. However, continuing with this line of thought, performing poorly on the SAT could also mean that the student has strong *unrealized* scholastic aptitude (or simply didn't put in enough effort on test day), so if the student were to study hard (not test prep *per se*, but school work) and take the test again they could do better.

While clearly preferable to many of the ideas associated with the early IQ testing movement, the ideas associated with aptitude testing bear resemblance to their predecessors in ways that remain problematic. Most importantly, aptitude and merit are thought to be solely within the individual, so little concern is given to the cultural surround or broader contextual factors that impact both test performance and broader psychological dispositions and traits. In Chapter 4, this neglect of culture and context was shown to be one of the essential problems with how the results of IQ tests were analyzed and interpreted. Secondly, aptitudes and merit are, like IQ, constructs that claim to explain a great deal about a person's abilities across a wide range of domains and into the future. While not as entirely encompassing of personal worth as the early IQ tests were taken to be, the SAT was (and often still is) taken as a general index of academic potential, which it is not and never was. So while claims to be measuring aptitude are an advance (both ethical and scientific) over claims to be measuring innate intelligence, they are still deeply flawed.

Nevertheless, Conant wanted aptitude testing, and he wanted it done with the utmost scientific rigor. To this end, ETS would innovate and pioneer new methods of test design, administration and scoring—constructing an historically unprecedented and highly reliable testing infrastructure of national scope. This was accomplished largely by leveraging emerging

computer technologies in the interest of securing reliability by way advances in "mechanical objectivity."

Reliability as mechanical objectivity

Under the leadership of Henry Chauncey, the president of ETS handpicked by Conant, and with the support of major philanthropic and federal influence, ETS quickly grew into a national institution (Lemann, 1999). This new era in testing rode a wave of advances in testing technology, including the invention of automated test scoring, as well as the first machine calculators (and eventually the first computers). ETS soon gained prominence as a high-tech, high-security symbol of American educational and scientific superiority (as a slew of 1950s *LIFE* magazine stories on ETS makes clear). The sheer quantity of data was unprecedented, and for a time in the 1960s ETS had a more sophisticated and secure computer system than the CIA, which made it perhaps the most complex in the world. ETS eventually came to attract the best minds in psychometrics as well as the support of a variety of powerful stakeholders, from military and intelligence agencies to the College Board, the NSF, a wide range of major universities, and a set of international scientific organizations. The SAT quickly became a staple in the lives of nearly all high school students. Other ETS tests soon became similarly taken for granted as necessary aspects of individuals' educational biographies, including the GRE (still used as a part of graduate school admissions) and the TOEFL (the most widely used test of English language proficiency).

The uniqueness of the metrological situation created by ETS cannot be stressed enough. Never before had a single test been given to all college-bound teenagers, many thousands of whom would all sit down at the same time on the same day at official and secure testing sites

throughout the country. Their answer sheets would be shipped across the country in trucks accompanied by police escorts and brought to the ETS campus in Princeton, New Jersey, where they would be feed into automated scoring systems. The sheer efficiency and scope of the biggest and most complex testing procedures in history bring to mind Porter's (1995) phrase "large-scale mechanical objectivity"— objectivity that is secured through techniques and instruments that assure vast numbers of subjects receive precisely identical treatment. Moreover, massive sample sizes and advances in quantitative psychometrics allowed for reliability and validity studies that further demonstrated the sophistication and rigor of ETS's "new science of mental measurement" (Lemann, 1999).

These technical innovations toward securing high levels of reliability are perhaps the most significant aspect of the revolution ETS initiated in the world of standardized testing. Recall that the early IQ testing movement created injustices right off the bat, simply by lacking reliability (e.g., carelessness in item design, inconsistencies in test administration, errors in scoring and statistical analyses). These kinds of basic forms of reliability are a condition for the possibility of any and all high-stakes testing, and it is a significant ethical advanced on the part of ETS in recognizing just how important reliability was in large-scale testing infrastructures. By prioritizing the pursuit of technical innovations in the interest of reliability, ETS solved many essential problems facing large-scale standardized testing, and indeed made some progress toward making justice-oriented testing a reality.

Thanks to these significant advances and their widely publicized scientific and academic cachet, it would not be until the 1970s that some of ETS's practices would come under widespread public scrutiny. Both affirmative action and the "Truth in Testing" bill posed complex and highly publicized legal battles (Nairn, 1976; Lemann, 1999; Spring, 1989). But by

this time ETS was serving a *seemingly necessary* public function. Having consolidated the majority of testing businesses and research entities, it was the premier provider of objective measurement to a diverse set of educational institutions. Moreover, ETS had the money to invest in lawyers and public relations campaigns, as well as in academic research favorable to the continued use of its tests. Seemingly above the fray and "too big to fail ([or] make a mistake, or be de-legitimated)," ETS remains today the largest and most prestigious testing organization in the world (Lemann, 1999).

The potential injustices of meritocracies

History shows, however, that some of Brigham's early concerns appear justified. As many commentators would point out: the commodification and large-scale institutionalization of a test makes it harder to change when research suggests it should be changed (Gitomer, 2010; Koretz, 2010; Lemann, 1999; Nairn, 1976). This dynamic is a clear example of the co-constitutive relations between testing and schooling discussed in Chapter 3. A test creates certain policies and cultural norms, which in turn reinforce the test. Changing an existing test is not just a matter of doing more research; it is a matter of altering existing cultural norms and social relations in the school system. The remarkable gains in objectivity that ETS accomplished due to its centralization of design, administration, and scoring were bought by sacrificing basic research into the relations between learning, testing, and educational practice. Granted, ETS continued to do research on the validity and reliability of its existing instruments and on the performance of new similarly designed instruments, but it avoided the kinds of research that could invalidate, overturn, or force a reconsideration of their fundamental ideas and practices, i.e., kind of

research that would constitute a true science of educational testing (Nairn, 1976; Lemann, 1999; Lagemann, 2000).

However, the main issue here is not the tradeoff between the benefits of broadly institutionalizing measures and the benefits of being able to easily and innovatively revise them. The real issue is the relation between the truncated objectivity of the SAT and the meritocratic ideal that motivated its use as the central component in a national objective testing infrastructure. As discussed in Chapter 2, the term *meritocracy* is used to refer to a system of institutions—typically but not always involving a society's educational system—that rewards individuals based on their possession of a certain set of skills and abilities. *Merit* is a constellation of traits deemed valuable in a certain socio-cultural context; building a social system that runs on rewarding the obtainment of said traits creates a meritocracy. Such a system is typically contrasted with an aristocracy, which rewards individuals based on their birth or familial lineage.

As explained below, meritocracies are *not by definition unjust*. In fact, in many contexts, including those in which ETS first thrived, a meritocracy that is built around mechanisms for objectively determining merit leads to an increase in social justice. By replacing an aristocratic college admissions system with a meritocratic one, the efforts of ETS represent some of the most remarkable efforts toward justice-oriented testing. However, as argued by Habermas (1984) and other social theorists, innovations designed in the name of justice often create new social justice issues through their very success (as advances in medical care ended up creating unprecedented and unanticipated social justice issues concerning the resulting health care system). This is just what happened with the test-based meritocracy created by ETS, which advanced social justice while at the same time creating new potentials for injustice due to the lingering concerns about objectivity, relevance, and benefits of its approach to testing.

In one of the most influential educational works from this era (or any era), *The Process of Education,* Jerome Bruner ambivalently endorses the then-emerging meritocracy. He gives voice to concerns about its impact on individuals while also arguing for its necessity, inevitability, and value—given the overall goals deemed critical to national defense during the Cold War:

> The present National Defense Education Act is only the beginning.... The peril of success under the conditions [it promotes] is the growth of what has been called "meritocracy." Partly out of inertia of present practice and partly an inevitable consequence of the national security crisis, there will be a strong tendency to move the able student ahead faster and particularly to move him ahead if he shows early promise in technical or scientific fields. Planned carefully, such acceleration can be good for the student and for the nation. A meritocracy, however, implies a system of competition in which students are moved ahead and given further opportunities on the basis of their achievement, with positions in later life increasingly and irreversibly determined by early school records. Not only later educational opportunities but subsequent job opportunities become increasingly fixed by earlier school performance. The late bloomer, the early rebel, the child from an educationally indifferent home—all of them, in a full-scale meritocracy, become victims of an often senseless and irreversible decision. A meritocracy is likely to have several undesirable effects on the climate in which education occurs, though with advanced planning we may be able to control them. One consequence may be an overemphasis upon examination performance.... Further, if the principle scholarships and prizes come increasingly to be awarded

> for merit in the sciences and mathematics, then we may also expect that there will be a devaluation of other forms of humanistic scholarly enterprise.... (1960, pp. 76-80)

Bruner notes several important issues that were merely on the horizon at the time of his writing. These issues would become increasingly central in the following decades. His ambivalence toward the emerging meritocracy reflects his sense that it was both inevitable and valuable as an aspect of national security, and yet it was also potentially damaging to the educational system and the lives of individuals. He shared the belief (also held by those within ETS) that insofar as scientists and engineers are needed to assure the continued security of the country, then meritocratic mechanisms must be implemented to facilitate a steady supply of them.

Bruner focuses on the potential damage done to individuals by pointing out that such a meritocracy would unfairly impact the life prospects of individuals who do not fit the meritorious mold. Bruner studiously avoids explicitly arguing that rewards would be distributed along socio-economic lines, but his reference to "the young rebel [and] the child from an educationally indifferent home" point in this direction.

Bruner continues, warning that the narrow focus of the emerging test-based meritocracy is also likely to lead to the devaluation of humanistic scholarship. This shows, as discussed in Chapter 3, the tendencies of testing-intensive policies to define the scope of what is educationally valuable—testing infrastructures *set the range of categories from which all may choose* through a kind of anonymous power (Busch, 2011). Bruner saw these significant dangers, but believed they could be avoided through proper planning—e.g., by assuring a lack of bias and

broadening the range of tested subjects. But these are not the core social justice issues raised by the institutionalization of a national test-based educational meritocracy. There are more fundamental issues that concern the relationship between testing and the very idea of meritocracy, which Bruner does not raise in the quote.

First consider the very idea of a meritocracy in the first place. Merit is, by definition, a set of skills and traits deemed valuable in a specific socio-cultural context; awarding merit is thus part of a broader incentivizing strategy toward pursuing the goods valued in that socio-cultural context. This a point clarified by Amartya Sen (2000, p.5), who following Rawls argues that "the concept of 'merit' is deeply contingent on our views of a good society. Indeed the notion of merit is fundamentally derivative... it is dependent on the concept of 'the good' in the relevant society." There is no "natural order of merit independent from our value systems" (*ibid.*, p. 10); theories of merit must therefore draw upon other normative theories. "The rewarding of merit is, to adapt a Kantian distinction, a 'hypothetical imperative' that is dependent on the way we judge the success of a society; it does not involve a 'categorical imperative' concerning what should in any case be done" (*ibid*. p. 14).

Recall the distinction between the reasonable and the rational that concluded Chapter 3. This distinction was said to retrofit the Kantian distinction just noted between hypothetical and categorical imperatives. Rationality implies only the intelligent pursuit of ends; reasonableness also implies the intelligent pursuit of ends *as well as* a willingness to abide by fair terms of cooperation with others. A meritocracy can thus be perfectly rational (e.g., scientists are needed, so intelligent means are devised to select and reward them), while also being entirely unreasonable (e.g., those deemed not fit to be scientists are systematically marginalized). Sen argues that meritocracies do not typically take into account the scope of inequalities resulting

from the reward of merit and therefore tend to create unacceptable inequalities of wealth and opportunity. This is why even a meritocracy based on a truly disinterested and objective mechanism for rewarding merit can still be profoundly unreasonable and unfair (the idea of an objective mechanism that can detect *merit* is further discussed below).

However, injustice does not necessarily follow from the reward of merit. It is possible to imagine a meritocracy where merit is defined so that social justice is pursued through its reward. Recall from Chapter 1 that Rawls's difference principle limits the shape and scope of all inequalities, even those that result from the reward of merit. Sen suggests that if such a principle were included in the overarching values of a society then merit itself would be recast in "an inequality-sensitive way.... [It is] the ad hoc exclusion of distributional concerns from the objective function in terms of which merit is characterized [that] makes meritocracies prone to generate economic inequalities" (*ibid*, p. 15). That is, if merit is defined in terms of a social ideal that *includes* considerations of distributive justice, then rewarding merit would mean rewarding those traits that lead individuals to advance a fair allocation of goods (e.g., perhaps traits like generosity, moral sensitivity, etc.). But when merit is defined in ways that *exclude* such considerations, the traits deemed valuable are likely to be ones that generate greater inequalities when they are differentially rewarded (e.g., analytical intelligence, extrinsic motivation, etc.)—creating a situation wherein those already advantaged receive additional advantages and the position of the least well-off is made worse. Unfortunately, all known large-scale educational meritocracies (including the one created by ETS) define merit in terms of a narrow social ideal that does not include social justice as the goal pursued through its reward (Arrow, Bowles, & Durlauf, 2000).

Moreover, Sen (2000) argues that meritocracies can negatively impact the self-understandings of those involved, potentially undermining their ability to understand themselves as free and equal citizens. There is a tendency for rewards based on merit to be understood as if they are somehow *owed* to the recipient, a tendency for those with certain traits to feel they *deserve* the special advantages they have been given through meritocratic mechanisms. Likewise, those not rewarded come to understand themselves as *not deserving* rewards, as somehow less worthy of social support and acclaim than others. Sen notes that this results from definitions of merit that exclude broader considerations of social justice, and which define merit in terms of what individuals can contribute to overall social efficiency (e.g., economic productivity; technical knowledge). This leads individuals to see their own self-worth in terms of their position in a hierarchy ranging from those at the top (who are capable of contributing to society) to those at the bottom (who are less productive and in many ways dependent upon the contributions of "their betters"). Moreover, because of the seeming objectivity of the mechanisms used to detect merit, an *inequality-insensitive* meritocracy can create the appearance that great discrepancies in wealth and opportunity are *legitimate and just*—systematically distorting everyone's sense of who is entitled to what and why.

These concerns about meritocracies are all reasonable and important, yet both Bruner and Sen do not raise what is, in fact, the essential issue: *how can merit be objectively identified*? During the first decades of its use the SAT was believed to measure "scholastic aptitude" and this was taken as a proxy for merit (Lemann, 1999; Narin, 1978). That is, doing well on the SAT was rewarded because it was understood as an index of academic merit—what the test measured was believed to be what the social system ought to reward. Assuming for the sake of argument that "scholastic aptitude" is what the SAT measures, there are still questions about if this

construct is appropriate as a proxy for merit. At the risk of being simplistic: why not creativity, ethical reasoning, or caregiving as a proxies for merit? Or if it is only a mater of academics: why not research skills, breadth of interest, or observational acuteness as proxies for merit? "Scholastic aptitude," as measured by a single test, is an unreasonably narrow and limiting way of defining merit, even in strictly academic contexts. As Sen notes, and Bruner fears, narrow conceptions of merit can distort the distribution of rewards (as only those who display a narrow set of traits receive them) and negatively impact self-understandings (as we internalize assumptions about our self-worth based on an unreasonably narrow set of evaluative categories).

Of course, the fact that the SAT does not measure "scholastic aptitude" makes these concerns mostly beside the point. What was intended to be a test targeting a broad psychological construct with a great deal of importance for an individual's life was slowly reduced to a test that is understood to be useful merely because of its predictive validity. Yet, the reward of merit is not typically understood as being based on predictions of success; merit is a trait worthy of reward whether success is likely or not. Moreover, the claim that someone has merit is a claim *about something*; reference is made to some set of traits possessed by the individual. As discussed above, the SAT is not strictly speaking about anything, except how well individuals do on the SAT, which is then shown to modestly correlate with college success. This reduces the argument that SAT performance should be rewarded to the idea that those likely to succeed should be rewarded. Furthermore, this reduces merit to a mere prediction of success in small range of academic contexts, a far cry from the substantive assessment of aptitude and merit Conant was after.

Recall Gould's (1996) reflections about the difference between *testing as a practice for setting limits* and *testing as a practice for enhancing potentials*, which were discussed in Chapter

4. Test-based meritocracies tend to enhance the potentials of those already advantaged and those who are already likely to succeed. At the same time, test-based meritocracies set limits on those who are less well-off. Instead of serving to enhance the potential of everyone (and especially the least well-off), testing comes to serve as a sorting mechanism that assures resources and opportunities are given to those who are understood, rightly or wrongly, as better able to make use of them.

If tests are only designed to function as sorting mechanisms then they can be built and administered in terms of a truncated notion of objectivity: they can be understood to "work" even though they are not about anything, usually because they have been shown to have some predictive validity. *This truncation of objectivity results in a lack of relevance and benefit.* Such tests can produce no knowledge as to how students doing poorly might be helped (i.e., not helped to do well on the test, which is easy, but helped to do well in college). Moreover, because such tests function as gatekeepers, and yet do not reference any clear skills or capacities, the only way to succeed is to focus entirely on the test itself. Working on developing the deeper capacities the test measures is a non-starter because these capacities (if they exist) are not clearly characterized or understood, and therefore test results cannot be translated into the kinds of pedagogical content knowledge needed to do the work. So the test itself becomes the center of pedagogical attention, which can lead to a valorization of test-prep as if it were actually educational. The fact that intensive test-prep itself can undermine the purposes and validity of testing practices is a topic that will be revisited shortly.

Tests, especially high-stakes ones, are not typically understood as providing insight that is pedagogically relevant and beneficial, and nor are they built to do so. As a result they drastically narrow and distort perceptions of educational value, draw an unhelpful amount of

attention to themselves, and lead to educational institutions that fall short of delivering on their commitments to social justice. At the same time ETS established the most reliable testing infrastructure in the world, and thus advanced the cause of justice-oriented testing, it normalized the truncation of objectivity and related inarticulacies about the educational meaning of test results, which has resulted in new insidious forms of test-based injustices.

Testing reified: the normalization metrological injustice

By centralizing research and development in the testing industry and institutionalizing a massive and highly reliable testing infrastructure (not to mention spending millions on public relations campaigns), ETS created an image of testing in the public imagination that placed its scientific acumen and social utility almost beyond question. The success of ETS has deeply impacted commonsense notions about the functions and purposes of educational measurement—depoliticizing testing infrastructures and allowing them to slip into the background, as measurement infrastructures tend to do—allowing a certain form of testing to become a taken-for-granted aspect of social life. The consequence has been, as just mentioned, the creation of an educational system centered around competitive performance on high-stakes tests, wherein the tests are presented as non-negotiable, apolitical, neutral scientific instruments. Thus those who "win" are understood as the rightful recipients of resources, opportunities, and acclaim—the tests scientifically prove they are deserving; for those who "lose," on the other hand, the tests scientifically prove the opposite.

Importantly, for those not favored by the meritocracy the test is perceived as a problem only in the sense that it is an obstacle; the ethical, political, and scientific status of the test itself are *rarely problematized*. Because of the taken-for-granted status of the test as a basic structure

in the educational system, the negative impacts of the test are understood as the result of an *individual's failing* or *inability*, not as the result of the testing infrastructure itself. This idea has been discussed in previous chapters, with reference to the sociologist C. Wright Mills (1956; see also: Busch, 2011; Tyack, 1974), who argued that institutional structures like tests become "naturalized," which turns public issues (having to do with the institutions that govern social life) into personal troubles (having to do with an individual's inability to succeed).

This is a crucial idea that challenges those who continue to promote testing as merely a scientific and technical undertaking, and thus neglect the profound political and ethical issues that are implicated in the design and implementation of testing infrastructures.

> By lending the rhetorical prestige of science to what may be questionable practices of an educational bureaucracy and a stratified economic system [i.e., by "disguising the politics of testing in terms that are merely scientific and technical"] there is no opportunity for rigorous attempts at examining institutional culpability.... Attention is primarily paid to [scientifically measuring] students' specific educational "problems," and thus, there is a strong inclination to divert attention both from the inadequacies of the educational institution itself and what bureaucratic, cultural, and economic conditions caused the necessity of applying these [measures] originally. (Apple, 2004, p. 127-128)

As discussed in Chapter 3, efficiency-oriented testing depoliticizes the social relations that result from testing by reframing them as technical problems (such as cost-benefit analysis and quality-control surveillance). Likewise, meritocracies justify social inequalities in terms of

similarly depoliticized objective mechanisms, with ramifications on how individuals understand themselves and their place in both the educational system and the broader social structure.

The meritocracy established by ETS is the first example of an efficiency-oriented testing infrastructure of national scope. Recall from Chapter 3, that efficiency-oriented testing, while providing some forms of objectivity, nevertheless overrides an individual's metrological rights in terms of relevance and benefit, and thus undermines the ability of educational institutions to provide for the allocation of primary goods, equality of opportunity, and self-actualization. All of this has proven true of the test-based meritocracy established by ETS. So while the creation of ETS was guided by a virtuous vision and admirably solved some ethical problems (such as replacing the prevailing aristocracy and amending the lack of reliability and scientific rigor in the testing industry), in so doing ETS spawned a new set of unprecedented ethical issues. Moreover, as will be revealed in the next section, ETS normalized the idea of reforming educational systems through the use of large-scale high-stakes testing, ostensibly backed by scientific expertise, and wielded in the service of narrow conceptions of social efficiency.

No Child Left Behind: the decline of objectivity and the inefficiency of injustice

> As 2014 neared, states were spending hundreds of millions of dollars each year on testing and on test prep materials; the schools in some districts and states were allocating 20 percent of the school year to preparing for state tests. This misallocations of scarce resources was hardly surprising, because schools lived or died depending on their test scores....This unnatural focus on testing produced

> perverse but predictable results: it narrowed the curriculum; many districts scaled back time for the arts, history, civics, physical education, science, foreign languages, and whatever was not tested. Cheating scandals occurred in Atlanta, Washington, D.C., and other districts. States like New York manipulated the passing score on state tests to inflate results.... Teaching to the test, once considered unprofessional and unethical, became common practice in the age of NCLB.
>
> -Diane Ravitch (2013, pp. 13-14)

Large-scale standardized testing in schools began during the first decades of the twentieth century as a quasi-scientific method for restructuring what were fast becoming America's most complex, crowded, and politically contentious public institutions. Mass-administered multiple-choice IQ testing swept the country as part of a broad social movement that promoted the *justice of efficiency*. Sorting students into groups according to test results was thought to maximize the efficiency of the school and prepare students to assume their appropriate place in the broader social system, thus also maximizing the overall efficiency of the economy. Justice was being administered, it was thought, because each individual was scientifically routed into their rightful place in the hierarchically structured social system; social harmony and optimal economic output would result—a vision explicitly reminiscent of Plato's *Republic*.

Yet these first testing infrastructures were wedded to racist ideology and biased by design, rendering moot questions about the justice of the society they were intended to create. Injustice was done to every child tested because of the profound lack of objectivity that characterized the instruments and procedures used. By midcentury this fervor for testing and its

accompanying ideologies had been scientifically sublimated into a single consolidated national organization, ETS, which would make unprecedented progress toward making objective educational measurement a reality. ETS changed public perceptions and common-sense notions about the function and value of testing, creating a test-based meritocracy surrounded by an aura of technological and scientific expertise and backed by every major power in the postwar educational establishment. A vision of social justice couched in terms of social efficiency remained the overarching motive, only now this vision had become wedded to an emerging science of education and increasingly powerful technologies for test design and administration.

During the first decades of the twenty-first century, large-scale standardized testing is once again transforming the educational system. It is expanding its role in schooling, such that in the coming years more students will be taking more standardized tests with greater consequences attached to them than at any other time in history, both absolutely (total numbers of tests taken; total amount of money and opportunities at stake) and relatively (number of tests taken per student; total amount at stake per student). This means that the impact of testing on the educational experiences and life prospects of individuals will be greater than ever before. Technological, political, and scientific factors have again converged to allow for a testing-enabled reformation of the educational system.

Playing an essential role in these developments are contemporary educational reformers who argue for the social justice benefits of testing. Today's testing-intensive reform efforts are championed and led by well-intentioned educational entrepreneurs, civil servants, and social scientists. Leaning heavily on arguments about efficiency and the science of education, a new vision of the relations between testing and social justice is emerging. These reformers argue that testing is an essential part of reforms that will transform our school systems, which are currently

unjust, inefficient, and anachronistic. They envision a brave new world of testing in which computer technology in the classroom will be leveraged to administer more and better tests, which will yield massive data-sets that can be used to determine teacher quality and school performance. This classroom technology will also be used to provide system-level overviews for researchers and policymakers. Testing is understood as a necessary component for reforms that will diversify and reinvigorate the educational landscape: tests are needed to quantify the value of new innovations, judge the payoff of educational entrepreneurship, exercise quality control at the level of the school and the district, and empower consumers as the educational system is transformed from a public bureaucracy into a complex marketplace. For the first time in history, US schools will be organized in terms of national standards and a related national testing infrastructure.

Unfortunately, these visions of testing seem to ignore the important ethical lessons that can be learned from even a cursory review of the history of testing, such as that presented so far in this work. In the decade or so since NCLB was signed into law—initiating this new era of testing-intensive reform—evidence has continued to mount that suggests history is repeating itself. There is evidence of a decline in objectivity reminiscent of the early days of IQ testing, which is being catalyzed by an extreme emphasis on efficiency as an institutional virtue. The basic idea of a test-based educational meritocracy has been expanded to include not only competition between students but also competition between teachers and schools. The overall vision of social efficiency puts perceived economic and political needs above the rights of students, teachers, and anyone else caught up in the dynamics of the rapidly transforming school system. Evidence suggests that by many standards American school system is quickly becoming the least equitable among industrialized nations, with some regions and socioeconomic sectors

offering what are among the best schools in the world, while schools in other areas can scarcely provide a safe physical environment for students, let alone educational opportunities (Kozol, 2012). Critics have argued for years that the current path of marketization and high-stakes testing is likely only to increase these inequalities, despite overly optimistic technocratic rhetoric to the contrary (Hursch, 2008; Apple, 2001; Ravitch, 2010; 2013).

In the absence of synoptic and well-established historical accounts (like those used in the previous historical sections), this section focuses instead on a limited set of well-documented recent occurrences. The discussion centers around accounts compiled by RAND (2010) and Ravitch (2010; 2013) that document the profound transformations in school cultures and practices that resulted from the institutionalization of federally mandated efficiency-oriented testing. Cheating was rampant and there were forms of intensive test-prep so extreme that they left the validity of the resulting test scores in question. Moreover, tests designed to reliably measure a limited range of student performance were repurposed to measure school quality, properties that these tests were never intended to measure (Koretz, 2008). These basic problems indicate a profound decline in concerns about objectivity. Beyond the truncation of objectivity discussed above, a situation emerged under NCLB where even basic reliability and validity were lacking in the testing practices undertaken in many schools. But even where testing practices were less obviously problematic, the effect of NCLB on school cultures was still devastating. *Curricula were drastically limited* to those subjects that would be tested. Teachers were systematically de-professionalized. Students were increasingly subject to distorted social relationships resulting from the test-based categories through which those relationships came to be filtered. The violation of the metrological rights of students and teachers and the resulting

inability of schools to offer what justice demands led to a widespread sense that NCLB was making schools worse places to learn and teach.

Overall, evidence suggests that the strategy of intensive efficiency-oriented testing backfired, failing to produce the improvements in student performance it was intended to promote, as scores on benchmark national and international tests (such as the NAEP & PISA) have shown declines overall, even as scores increased on state-level tests (see: RAND, 2010; Ravitch, 2013). This leads to important arguments concerning the *inefficiency of injustice*: when the quest for efficiency becomes so great that justice is sacrificed, efficiency itself is then compromised by the social and cultural dysfunctions accompanying widespread injustice. This is considered in terms of the massive waste of money, time, and energy—not to mention the lack of learning and loss of educational value—that results from large-scale organized cheating.

NCLB: a recipe for injustice

Even a cursory review of the structures put in place by NCLB reveals that they were far from a neutral or beneficent force in schools, especially struggling ones. NCLB was complex and contained a wide variety of programs and sub-programs, but its main impetus was to put in place a form of efficiency-oriented testing that included the following features (based on Hess & Petrilli, 2006):

- States were expected to choose their own tests and the performance levels on them that would define "proficiency."
- All schools were required to test all students in reading and mathematics at regular intervals from third grade through high school.

- All states were required to establish timelines showing how *one hundred percent* of their students would reach proficiency by 2013-2014.
- All schools were required to make annual yearly progress (AYP) toward the goal of *one hundred percent proficiency*.
- Any school not making adequate progress would be labeled a school in need of improvement (SINI) and face increasingly onerous sanctions, including being shut down and "restructured."

4) Schools that were required to restructure had a limited set of options, including: converting to a charter school; replacing the principal and staff; relinquishing control to private management; or turning over control of the school to the state.

In light of the discussions in previous chapters it should be clear that there would be significant problems with a testing infrastructure built and institutionalized according to these policies. First off, the system lacks objectivity as a result of each state being given the freedom to not only determine what tests are used but also what counts as proficient. The extent of this lack of objectivity will be discussed below. As discussed in the Conclusion, the new Common Core Standards and Assessments (CCS&A) seek to remedy exactly this problem, by providing a single testing infrastructure for all schools. This is a positive move in terms of the first principle of just institutionalized measurement. But a lack of objectivity was only one of NCLB's ethical failings, and it is likely that any testing infrastructure coupled to a set of punitive and unreasonable high-stakes polices will do damage, no matter how well built and objective it is.

As the policies reviewed above indicated, NCLB was based on the idea that *one hundred percent* of students could be made proficient by 2014 (13 years from the date it was signed into

law). Critics (Hess & Finn, 2007) have lampooned this as not only unreasonable for sociological reasons (e.g., seeing it as analogous to the idea of cities being one hundred percent crime-free) but also as a mathematical impossibility that ignores the basic laws of statistics, such as the ubiquity of normal distributions and regressions to the mean (e.g., seeing it as a logical error akin to the so-called Lake Woebegone effect, a community in which "every child is above average"). Moreover, every school was expected to make consistent linear annual progress toward the goal of one hundred percent proficiency *or be punished*. While the rhetoric was that failing schools would be helped, they would in fact be shamed and shut down and new schools would be put in their place. Teachers and principals would lose their jobs and a new (often privately run) school would replace a public school that had (with its name, sports teams, and traditions) been a part of a community for decades.

This is the second basic ethical failure of NCLB: systems that further disadvantage those who are already disadvantaged are unjust. These are echoes of the problems with test-based meritocracies discussed above. This was an injustice built into the basic design of NCLB as a law, which was combined with a lack of objectivity and the setting of unreasonable goals (one hundred percent proficiency). It was a recipe for injustice.

> Because NCLB required states to promise that they will reach an impossible goal, the states adopted timetables agreeing to do what they couldn't, no matter how hard teachers and principals try....With every passing year more and more public schools failed to make AYP and were labels as "failing"... even though some states lowered the cut scores (or passing marks) on their tests to make it easer for schools to meet their target.... In the school year 2006-2207, 25,000 schools did

> not make AYP. In 2007-2008, the number grew to nearly 30,000, or 35.6 percent of all public schools.... The consequence of mandating an unattainable goal was to undermine states that had been doing a reasonably good job and to produce a compliance-driven regimen that recreated the very pathologies it was intended to solve. (Ravitch, 2010 pp. 103-104)

The decline of objectivity

One of these pathologies was the wide spread cheating that took place under NCLB. In the Atlanta Public Schools, for example, a system-wide cheating ring was organized from the highest levels of the administration down, involving hundreds of teachers, and the impacting the test results of thousands of students. Millions of dollars were involved and nearly a dozen individuals, mostly teachers and administrators, were arrested and will stand trial on racketeering and conspiracy charges, facing jail time for cheating on tests (Georgia Bureau of Investigation, 2011). This kind of highly organized cheating was not the norm, although other cities have had major issues, including Baltimore, New York, and Washington DC. Yet even when no explicit cheating was taking place, classrooms and school cultures were being negatively transformed due to the high-stakes environment created by NCLB. The biggest and most obvious concerns have to do with the ways in which the testing infrastructure resulted in narrowing the curriculum and increasing the prevalence of test-prep style instruction. These trends have been widely documented (Hursh, 2008; Koretz, 2008; RAND, 2010; Ravitch, 2010; 2013).

The place to begin, however, is with the decline of objectivity initiated by the very details of the policy itself. Recall that the law states that each state is responsible for its own testing infrastructure i.e., deciding on the test (but only from those for reading and math), the test

provider, the administration schedule, and most importantly, the proficiency levels or "cut scores" marking proficiency. The result was that definitions of proficiency "varied wildly from state to state, with 'passing scores' ranging from the 6th to the 77th percentile…. The testing enterprise [under NCLB was] unbelievably slipshod. It is not just that results varied, but that they varied almost randomly, erratically, from place to place and grade to grade and year to year in ways that have little to nothing to do with true differences in pupil achievement....The testing infrastructure on which so many school reform efforts rest, and in which so much confidence has been vested, is unreliable—at best" (Finn & Petrilli, 2007). Needless to say, this echoes many of the same issues that were raised surrounding the lack of basic reliability that plagued the early IQ testing movement.

Moreover, the tests deployed to evaluate school quality were tests designed only for measuring a limited range of student performances in math and reading (Koretz, 2008). Repurposing tests in this way is not necessarily problematic, especially when they are used in low-stakes contexts and opportunities are made for researching the reliability and validity of their new purpose. But this was not the case under NCLB. Tests were repurposed and immediately used for making high stakes decisions. It is understandable why tests of student performance might appear appropriate for measuring school quality: student performances can be thought of as the "outputs" of schools, and the better the "outputs" the better the school. However, this idea that measuring the "outputs" of schools is enough to tell you about its quality is based on the false assumption that all schools have comparable "inputs." Admitting that different schools deal with very different kinds of students undermines the validity of inferences about school quality based on measuring school "output." To use a simple analogy, it is like comparing the quality of two surgeons by using their death-rate statistics, only one is a plastic

surgeon and the other is a brain surgeon. It is simply not a valid comparison; the one will always do worse than the other by this measure because they are dealing with very different "inputs." This is not the place to detail the complexities of measuring school quality (see: *Ibid*). The point here is only that the repurposing of tests from measuring student performance to measuring school quality is problematic in terms of objectivity: tests about one thing are used as if they are about something else.

These practices that were built right into the law—variability between states, invalid repurposing of tests—reflect a serious decline in concerns about objectivity, which resulted in an almost obsessive emphasis on preparing for specific tests. NCLB resulted in a distortion of educational value like that described surrounding the SAT, wherein test-prep comes to be valorized as if it were educational. Many schools began to simply cut "non-essential" programs such as art, music, and physical education. Schools also cut back on "essential" programs like science, history, and civics. A nationally representative study conducted by the Center on Education Policy (McMurrer, 2007; 2008) revealed that 44 percent of schools reported a substantial reduction of time spent on science, social studies, and the arts, while 62 percent increased time on reading and mathematics. In many schools (again, especially those already disadvantaged) the months leading up to testing events involved the elimination of *everything* that was not going to be on the reading and English-language arts tests mandated by the state (Montefinise, 2007).

These same studies show that instructional techniques themselves were altered and constrained by the demands of the tests, devolving into the explicit teaching of test-prep strategies and repeated exercises involving practice problems, out of context, and without attention to the meaning of the content. As one New York City teacher told a reporter (quoted in

Ravitch, 2010 p. 108): "My students don't know who the president was during the Civil War, but they can tell you how to eliminate answers on a multiple-choice test. And as long as our test scores are up, everyone is happy. [Test scores] are our priority. Actual education is second." The sentiments expressed by this teacher are reflected in an increasing number of studies that show not only a massive increase in the money spent on test-prep materials and services, but also a clear transformation of teaching practice toward the forms of didacticism that mirror the limited content and format of the tests (Koretz, 2008; RAND, 2010).

It is important to understand how this kind of intensive test-prep can undermine the validity of the whole testing enterprise (Koretz, 2008). The vast majority of tests are not intended to be about how well students have been taught to take the test or how much they have focused on what will be on the test. Tests are typically intended to be about how well students know certain domains of knowledge or can exercise certain capabilities, *in general and on the whole*. Testing is meant to *sample* the quality of student capabilities, not become the sole focus of student learning. Tests lose their intended meaning when they become the sole focus of instruction; the claims we want to make using the results they yield are no longer valid. For example, take a simple test of academic vocabulary. It is intended to be a general index of vocabulary; so forty words are randomly selected from the thousands that constitute the domain. If a student does well with this random set of forty, then the inference that they also have a good grasp of the rest of the domain is justified. However, if the student is drilled precisely on the forty vocab words that will be on the test their success on the test can be no longer taken as an index of their facility with the broader domain of academic vocabulary. This could be shown easily by giving the student a different test containing a random forty words about which they were not primed. Their score on this test would be very different (presumably much lower) than

their score on the test for which they were so intensely prepared. By focusing only on the content of the test, intensive test-prep undermines the validity of the test as an index of the broader domain. This phenomenon is often referred to in terms of "inflated test scores" and is a serious but under-researched problem that undermines the validity of many forms of high-stake testing (*Ibid)*.

The reasons why schools and teachers would adopt such intensive test-prep strategies should be clear given everything said thus far about the impacts of testing infrastructures on schools. As explained in the discussions of the education commodity proposition and efficiency-oriented testing in Chapter 3, when test scores are used as the sole proxy for the value of educational processes it distorts perceptions of value, often limiting what counts as valuable education to only that which can be measured. The effects of this simple, powerful, but partial way of thinking are far reaching. Testing infrastructures set the terms by which choices are made in schools, setting the space of possibility for both pedagogy and learning. Efficiency-oriented infrastructures constrain education to what is measured on a narrow range of tests and will often lead to violations of the second and third principles of just institutionalized measurement. These testing infrastructures become insensitive and unresponsive to the needs of those most affected, bringing them no direct benefits and often causing harm. Violating the metrological rights of students and teachers creates educational environments that are unable to provide for the educational primary goods that are each student's right, and that undermine equality of opportunity and possibilities for self-actualization.

The inefficiencies of injustice

Because of the injustices it creates, efficiency-oriented testing infrastructures can counterintuitively generate greater organizational *in*efficiencies, typically as a result of escalating *surveillance and enforcement costs* (Bowels & Gintis, 1986). These costs are incurred by an organization when its members view the overall organizational structure as being an obstacle or in opposition to their interests, or when any member is systematically disincentivized from performing their role in it. This was mentioned during the discussion of efficiency-oriented testing in Chapter 3, where we saw that in general across industries quality-control measurement infrastructures tend to be expensive to build, maintain, and implement. This inevitably leads to concerns about the *cost of surveillance* needed to exercise certain kinds of quality control. In this case, NCLB can be understood to have created an educational environment in which teachers and students were the subject of quality-control surveillance on a massive scale.

The costs of surveillance are not only financial (*ibid.*). There are significant impacts on organizational cultures and individuals' self-understandings when surveillance (and related methods of enforcement) become necessary aspects of organizational functioning. This leads to organizations in which the reason individuals do their jobs a certain way is *not* because they agree with the effectiveness of the technique and appropriateness of the task, but rather because the "quality" of their work is closely monitored and deviations from the mandated methods are punished. If they were not so strictly and insinuatingly surveilled they would prefer to do things differently—*not* out of laziness or incompetence, but because they believe there really is a better way. The result is that less work gets done at a lower quality accompanied by steadily increasing costs of surveillance, employee turnover, and the distractions and low-morale impact of enforcement practices, not to mention an increasing probability for active expressions of worker

discontent, such as sabotage (e.g., as seen above, cheating under NCLB could be understood as a kind of sabotage). Bowels and Gintis (1998, p. 6) speak to the literal price (in dollars) of injustice:

> Institutions supporting high levels of inequality [and injustice] are often costly to maintain. [There is] a cost in enforcing inequality, in such forms as high levels of expenditure on work supervision and security.... [But there is a] positive relationship between efficiency and equality in that more equal societies may be capable of supporting [higher] levels of cooperation and trust.... Cooperation and trust are essential to performance [and efficiency]. Of course trust and cooperation do not appear in conventional economic theory.

Costs of surveillance go up more if the objectivity of the instruments used is hard to maintain. Poorly built equipment or the likelihood of human error (or deception) requires that quality testing be done under more strict and exacting conditions, which are more expensive both financially and psychologically. Costs also go up faster in industries where the 'output' being measured is not a simple object (like grain or a car), which can be tested and measured by means of uncontroversial physical instruments. In so-called service industries (as some would have education become), where the product is more intangible, quality-control monitoring is not so easy and is often much more invasive, subjective, and expensive (*ibid.*). Moreover, as mentioned in Chapter 3, there is always a tradeoff between the damage done to the product by testing it and the gains in quality that can be made through increasing the same process (Busch, 2011). Apples must be tasted, fuel burned, and drugs tested in order to determine and improve their quality. The

more you test something the better sense you will have of its quality and of how to improve it, but in doing so you will also have destroyed more of the product. While the product begins to lose value through more and more testing, this process costs more and more money. All of these lessons about the dynamics of institutionalized measurement apply in thinking about NCLB, in which testing as quality-control surveillance was exercised in blanket fashion throughout the nation's educational system.

As discussed in the Conclusion, the latest testing infrastructure being built as part of Obama's CCS&A continues what has been a more general trend of investing in surveillance technology. 'Improved test security' is a major leg of the argument for investing at the federal and state levels in an entirely computerized testing infrastructure modeled on the security platforms pioneered by ETS and its test security and computer center subsidiaries. It is not a coincidence that these new high-tech tests will make it impossible for teachers to get their hands on students' answer sheets at exactly the time when these answers will begin being used to officially determine each teacher's value-added quality as part of expanding test-based accountability. While there is a small countervailing discourse about the need to include teachers in building and evolving assessment practices (Apple, 2013), the general trend is quite the contrary. The systemic disempowerment and de-professionalizing of teachers is understood as part of an effort to improve the quality of their practices. This is the kind of theory-versus-practice inconsistency that marks an institutional configuration as unstable and crisis-prone (Bhaskar, 2013).

No doubt, it is important to monitor the quality of the educational processes that take place in schools, to ask questions such as: How good are the teachers in this school? How much has this child learned? This is essential. But the use of testing infrastructures as the dominant

index of quality leads to a distortion of value. This is a distortion in the meaning of what counts as a good education, and it creates a new ideal of what teachers and students ought to be and do. Testing can distort the perception of value to such an extent that 'quality control' becomes a counterproductive undertaking.

Attempts to 'steer' a complex system (such as a school system) typically fail when they are undertaken by focusing narrowly on one aspect of the system; 'steering' based on feedback tracking only a *true but partial* representation is bound to fail (Buck & Villines, 2007). When the system being steered is one constituted by complex human relationships (like those between teachers and students), a narrow measure of quality control will distort these relationships, leading to an increasing sense of injustice. False measures engender false-consciousness, disingenuousness, and systematically distorted communication (in the fully Habermasian sense of this term); or they occasion widespread discontent, disruption, push-back, subversion, and revolt, as discussed in Chapter 1 and as witnessed in the Atlanta Public Schools.

In the long run, 'steering' blind (or according to limited and misleading measures) will create situations in which even the narrow qualities that are officially and objectively monitored start to decline. This decline in 'official efficiency' is in fact a result of long-standing systemic disruptions in the culture and social relations of the organization, disruptions that were initiated as part of a misguided attempt to improve these very qualities. Something like this was discovered to be the case under NCLB, a trend shown clearly in states where the NAEP showed declining reading and math scores while state test scores showed steady gains in these areas (Ravitch, 2013).

So it is with the *inefficiencies of injustice* when the impact of injustices that result from a policy undermine the goals which that very policy was initially meant to achieve. Recall the

discussion from Chapter 1 about the unjust bureaucracy. Efficiency experts in agriculture bought (or seized) vast tracts of land, precisely measured and parceled out lots, set metrics for production, gave the occupying rural peasants heavy machinery (mostly unfit for their local conditions), and then asked them to more than quadruple their output. These modern efficiency techniques led to peasant revolts, equipment damages, and widespread crop failures and famine. They created conditions far worse than those experienced before the "scientific improvement" of what were ancient practices.

The *inefficiencies of injustice* are a common and problematic pattern that beset many modernizing practices (Bowls & Gintis, 1998; Porter, 1995), especially authoritarian forms of modernization (Scott, 1998; Apple, 2001). NCLB was beset by these inefficiencies stemming from injustice. Cheating, test-prep pedagogy, and rising costs of surveillance ultimately resulted in the 'kickback' of injustice's inefficiencies: increasing social inefficiencies (resulting from a focus on economic efficiency), worsening actual quality (resulting from increased focus on test-based instruction), and lessening and lax standards (resulting from increased standardization). Some of these problems can be resolved by improving testing practices—making better and more secure tests across a greater range of subjects. And this is the direction in which testing is headed. Once again bolstered by technological advances and opportune political climates, testing infrastructures are expanding and increasing their size, scope, and significance.

Conclusion: prelude to the future of testing

Before looking to the future of testing it is worth briefly summarizing the results of this chapter. ETS took a major step in the direction of justice-oriented testing by pioneering technical innovations and building a testing infrastructure of both unprecedented size and reliability.

However, while securing some forms of objectivity, this new approach to testing neglected others: what the tests were about became less important than their functional fit for institutional purposes, which came to be understood mainly in terms of predictive validity. This truncation of objectivity created situations in which the test unduly narrowed perceptions of educational value, thus leading to an unjust meritocracy based on an unreasonably narrow and increasingly meaningless index of merit. NCLB inherited this truncated form of objectivity and led, in turn, to an even greater decline of concerns about objectivity. Reliability fell back below the levels established by ETS and forms of intensive test-prep largely invalidated many of the testing efforts. Large-scale efficiency oriented testing resulted ultimately in a wide range of inefficiencies stemming from injustice, as cultural and pedagogical dysfunctions resulting from unreasonably narrow and overtly fetishized forms of testing began to undermine the goals for which testing was institutionalized in the first place.

Conclusion: social justice and the future of testing

> Part of the pressure for [efficiency-oriented testing] policies comes from educational managers in bureaucratic offices who fully believe that such control is warranted and "good." Tighter control, high-stakes testing and (reductive) accountability methods provide more dynamic roles for such managers.... Their own mobility *depends* on the expansion of both their expertise and the professional ideologies of control, measurement, and efficiency that accompany them.... The policies enable such actors to engage in a moral crusade and enhance the status of their own expertise.
>
> -Michael W. Apple (2001, p. 58)

> People say, 'Well, you know, test scores don't take into account creativity and the love of learning.' I'm like, 'You know what? I don't give a crap.' Don't get me wrong. Creativity is good and whatever. But if the children [can't pass the test], I don't care how creative you are. You're not doing your job.
>
> -Michelle Rhee, former superintendent of the Washington D.C. schools
>
> (quoted in a TIME magazine cover story, 2008)

Previous chapters demonstrated that efficiency-oriented testing has been the dominant framework in US education for the institutionalization of large-scale testing infrastructures. Testing has served a wide range of efficiency-oriented functions in schools, with an emphasis on meritocratic ranking, quality-control surveillance (accountability), cost-benefit analysis, tracking,

and sorting. The social pathologies and injustices resulting from efficiency-oriented testing have already been reviewed: pretensions of objectivity; justice-insensitive meritocracies; inefficiencies of injustice; distortions of value resulting from the terms of the education commodity proposition. These are well-documented, negative consequences of testing-intensive educational reforms. Yet these kinds of reforms, and the reformers who have carried them out, have almost always been guided by a vision of justice-oriented testing. Testing has always been inspired by ideals of social justice and supporters of testing have sincerely intended to do good.

As mentioned in the Introduction, (and to use the kind of medical analogies popular within the testing discourse), the injustices stemming from testing can be understood as *iatrogenic effects*. *Iatros*, from the Greek, means '*healer*,' and genesis of course means 'creation' or 'origin'; and so iatrogenic effects are illnesses that stem from medical treatment—preventable harm resulting from the actions of a well-meaning medical professional. Iatrogenesis occurs, if you will, when the medicine is worse than the disease. In our educational system, then, why have prescriptions for testing brought with them so many negative side effects?

For starters there have been flaws in many of the *ideals of justice* that guided the testing-intensive reform efforts of the past. The most obvious is the conflation of efficiency and justice, which has been discussed at length. This error typically appears as part of a belief system in which increases in system efficiencies are understood as increases in social justice. The idea is that efficient organizations maximize the creation of certain goods (from social goods like math skills to physical goods like cars), and because these goods benefit individuals (they are 'utilities') the efficiency with which goods are created and distributed can serve a proxy for justice. Recall the Roosevelt quote from Chapter 3, in which he extolled efficiency as a virtue, citing the preservation and right use of national resources as a kind of civic duty. The more

efficiently a society operates, the more goods that society produces, the greater the aggregate benefits that accrue. By this reasoning, efficiency is not merely a prerequisite for justice (which is true to a point, as will be discussed below); it is understood as *equivalent to* justice or as justice's premier means of implementation and manifestation.

There is a moment of truth in this idea, which is why it remains so prominent and continues to occupy legal decision-making and economic policy (Rogers, 2012). Systems that utterly lack efficiency, objectivity, logical consistency, and planning are wasteful and usually unjust—so are systems that appear to have these qualities but are error-prone or undermined by dishonesty. When starting in such a radically inefficient state, a little efficiency begets a lot of justice (like cases can be treated alike, records are kept, metrics are set, etc.); but too much emphasis on efficiency begets injustice. Beyond a certain point, efficiency-oriented practices begin to undermine justice as prior chapters have shown with testing.

Alternatively, the Rawls-inspired version explored so far differentiates justice from efficiency, clearly prioritizes the former over the latter, and articulates their interactions and mutual impacts. Justice has priority over efficiency as an institutional virtue, and this is only in part because justice enables efficiency. Justice deals with ends-in-themselves, whereas efficiency deals only in the means-to-ends; efficiency therefore ought to be subordinate to justice. When priorities are reversed, both suffer: overvaluing efficiency tends to create injustices, which then in turn tend to undermine efficiency. That is, when efficiency trumps justice the result is injustice, and the result of injustice is inefficiency. This is the 'kickback' that results from an overemphasis on efficiency, discussed above as the *inefficiencies of injustice*. When this lesson is learned, future testing infrastructures will be designed to prioritize social justice, and will perceive gains in efficiency as collateral benefits resulting from improvements in social justice.

This is perhaps one of the most powerful insights yielded by this work: injustice is inefficient, while justice promotes efficiency. This may seem like a counterintuitive idea given the popular characterization of justice and efficiency as conflicting institutional virtues and the common idea that justice must be self-sacrificial and opposed to things like profit. But the notion that justice actually promotes efficiency is an idea that has been around for a while, finding support not only with Rawls (1971), but with Habermas (1984), Sen (1991), Bowles & Gintis (1996), and Nussbaum (2001). Efficiency and justice need one another; they are mutually supporting goods. Justice-oriented testing transcends but includes efficiency-oriented testing, constraining and focusing the use of testing practices so that *efficiency serves justice*. Yet what exactly does this mean and how can it be institutionalized?

This is the question addressed here in the Conclusion, a discussion that is admittedly preliminary and speculative. Justice-oriented testing must leverage advances in educational technology and the learning sciences (e.g., psychology, neuroscience, learning theory) in order to design testing infrastructures that are able to meet the needs of students and teachers first. The first section below looks at emerging trends in educational technologies that will be having an impact on testing infrastructures in the coming decades. The following sections suggests how these new technologies might be informed by the learning sciences, offering a set of design principles for justice-oriented testing. The final section is the most speculative, focusing on the need to reorganize the social relations of the school in order to assure the continued advance of justice-oriented testing. Being committed to designing tests that benefit students and teachers also requires putting in place broader institutional decision-making policies that include and foreground the experiences and values of students and teachers (i.e., not merely using these individuals to gather data for use in organizational decision-making process in which they have

no voice). Taken together, advances in educational technology, the new sciences of learning, and the politics of school cultures will have an impact on the future of testing — for better or for worse. As such, the goal of this conclusion is simply to suggest preferable directions for these essential aspects of future testing infrastructures.

Educational technology and the future of testing

> [In 2012] the U.S. Department of Education awarded 350 million [dollars] to two consortia to develop national assessments to measure the new national standards..... State education departments warned that the enhanced rigor of the Common Core would cause tests scores to plummet by as much as 30 precent, even in successful districts.... This, in turn will create a burgeoning market for new products and technologies.... This burst of entrepreneurial activity was planned.... Race to the Top was designed to scale up entrepreneurial activity, to encourage the creation of new markets for both for-profit and non-profit investors. The new standards [and tests] were a linchpin to match "smart capital" to educational innovation.
>
> -Diane Ravitch (2013, p.16)

The Department of Education has recently funded two large test-development consortia tasked with building the beginnings of a new national testing infrastructure, which is to be wedded to the new Common Core Standards (US department of education, 2011). The call was for new assessments that measure twenty-first century skills and go beyond multiple-choice

items and shallow assessments of content. While the shape of tomorrow's national testing infrastructure is still indeterminate, we can be sure it will be larger, more complex, and increasingly integrated with new technologies. A few welcome advances should help with some of the social justice issues raised above, such as lack of objectivity, which was an explicit concern in the discourse surrounding the dismantling of NCLB and its replacement by the CCS&A.

Some researchers have argued that irrespective of prior *testing policies*, fundamental changes in *testing technologies* have rendered all prior testing infrastructures obsolete (Collins & Halverson, 2009). A report from the National Science Foundation task force on cyber-learning (NSF, 2008), focusing on the future of educational technology, describes burgeoning markets for educational technologies that rely on high-speed Internet and powerful computing. Postindustrial-era testing infrastructures will be built upon whatever new technologies come to play a dominant role in education. The NSF suggests that testing will likely become wedded to technologies that allow for embedded testing that is repeated, formative, and in the service of real-time learning (*ibid*.).

There is a great deal of speculation concerning tomorrow's educational technologies—the so-called emerging "edu-tech." In particular, there are some important characteristics and trends in emerging educational technologies that are relevant to testing. The recently funded consortia tasked with building the first Common Core Assessments are only an initial attempt at building "next generation" testing infrastructures of national scope. It is impossible to know what testing will look like in the coming decades, but we can think through possible futures based on the trajectories of major trends characterizing emerging educational technologies (based on: Collins & Halverson, 2009; NSF task force on cyberlearning, 2008; Dawson & Stein, 2011):

Technology saturation: Computers and other networked devices will reach an increasingly large proportion of the population, especially students. Smart phones and other portable devices are already ubiquitous.

Just in time learning: Learners will use technologies that organize databases of resources that allow for instant access to whatever needs to be learned.

Customization: Learners will use technologies that are responsive to individual differences as educational opportunities are guided by user preference and performance.

Scaffolding: Learners will use technologies that structure the delivery of tasks and learning opportunities based on close-to-real-time assessments of performance.

Reflection: Learners will use technologies that document the history of user performances and then present comparisons among users' histories. Technology will enable detailed portfolio management systems, templates, scoring interfaces, and databasing.

Distance learning and online education: More education will take place at a distance through online learning environments, with improved efficacy due to advances in video conferencing and content-delivery systems.

Databases and electronic learning records: With embedded assessments and automated progress- and behavior-monitoring technologies, educational record keeping will become as detailed and complex as medical record keeping.

This list of trends and characteristics is not exhaustive, but it is nevertheless suggestive of what tomorrow's educational institutions will have to work with. The history of testing teaches that advances in technology can radically alter testing practices and in turn radically alter the structure of the educational system. Advances in testing technology have always forced educators to reconsider the function of assessment within the system. If this list captures the characteristics of the technologies that will become essential to future educational configurations, then the task here is to begin inquiring into the preferable directions for testing that they can make possible. Of course, the future of justice-oriented testing involves more than leveraging advances in technology. The focus here is to close out the critical historical narrative by venturing to write a 'history of the present,' and to foreshadow imminent problems stemming from the current trajectory of testing innovation.

The plans, sample questions, and press releases put out by the two CCS&A test-building consortia show that they are aiming to leverage many of the technological trends listed above, including customization, scaffolding, and large-scale databasing (Smarter Balance Assessment Consortium, 2014). All of the CCS&A tests will be administered via the Internet and taken at computer terminals or laptops. Many schools do not have the computers needed or have to make upgrades to their existing systems. The cost of ongoing tech-infrastructure upgrades alone, which are primarily being saddled upon each state, will be a large part of education budgets moving forward, and most of it will ultimately go into the hands of a few large technology corporations that sell to schools.

These tests will make use of keyboard skills, mouse skills, as well as some limited web browsing. They will provide questions that are customized to a student's proficiency level by delivering questions based on automated analyses of prior performance, a process ETS has used

for years and which can only be done on computer-administered tests. There will also be more tests, including so-called formative assessments, which will put students in front of computers to take sample questions and receive guidance on how to prepare to take the main test later in the year. More subjects will (eventually) be tested and a wider variety of item formats will be used, including the short answer, and a broad category known as 'computer-enabled selected response.' These 'selected-response' questions involve complex answer-selection tasks. For example, a student might highlight a section of text from an excerpt of a book as a way of answering a reading comprehension question: "Use your cursor to show where the character in the story expresses fear."

These tests admirably transcend simple multiple-choice item formats and also likely offer improved objectivity and security. But when pushed CCS&A test designers rightly admit that the new tests are largely exploratory and that a great deal of work remains to be done (SBAC, 2013). The first of these next-generation tests will be rolled out in 2014, but iterations and revisions will follow, as more is learned about what it means to run a testing infrastructure of such unprecedented complexity, cost, and significance.

The future of testing: more, better, faster, now

Even just to ensure the objectivity, reliability, and validity of new testing technologies a great deal of research is necessary. These areas of research are simple, but unexplored: how do computer skills differ by SES and what is their impact on test scores? As tests expand to laptops and tablets, are there are differences in performance across different hardware and software interfaces? The effects on students of more numerous and extended forms of testing as well as the impact on pedagogy and curricula of testing across more subjects also remain to be seen.

While the medium of administration is different, a high-stakes test is a high-stakes test—it will impact the classroom, the teacher, and every student accordingly.

Moreover, while 'selected-response' items (with dozens if not hundreds of possible answers) are more complex and challenging than traditional multiple-choice questions, they remain a staple in the new computer tests only because they are amenable to automated scoring. They are not open-ended nor do they elicit expressive or creative student responses. They also contribute to making the experience of testing into one where individuals feel alienated and powerless. As if nodding to B.F. Skinner's learning machines (and taking Porter's notion of 'mechanical objectivity' literally), computerized selected-response tests are experienced by-test takers as "like being evaluated by a machine." This uncanny experience (of having critical life-prospects determined by seemingly mechanical means) obscures the fact that educational testing is always an interpersonal occurrence. The computer is enabling a 'conversation' (albeit one way) between the student and a wide range of interested parties, from their teacher and local school administrators, to the state and national DOE, as well as research agencies, think-tanks, and philanthropic investors.

Because of the perceived value of testing, and the promises of new technologies, there are initiatives to design new assessments that focus on traditionally non-tested subjects, like art, science, and history. This expansion of assessment areas will be coupled with the ongoing evolution of the "computer-technology education complex," sketched in the tech-trends listed above. Testing will be expanded as technology continues to evolve its role in schools. Testing will diversify across platforms as it rides the educational impacts of the tech-trends listed above. Even a quick look at this list of emerging trends shows that as testing diversifies across computer

platforms the themes of surveillance, privacy, and efficiency will remain pressing ethical concerns.

Video games and social networking sites capture thousands of data points per session about their users, all housed in a central data storage system that can be analyzed and searched (Pariaser, 2011). Educational assessment technologies are still catching up in the capture of real-time computer-enabled behavior-tracking. As teaching and learning move from pen, book, and paper to glowing screens and operating systems, educators can now make use of large-scale "learning-analytic" back ends built into their e-learning platforms. This allows for cataloguing student scores on practice tests and official tests, as well as what was read, googled, or viewed, for how long, and where. Many schools that own the computers used by students withhold students' rights to privacy and allow administrators to remotely view student computer use, including webcams (Spring, 2010). Embedded assessment takes on a whole new meaning when testing technology is literally embedded in students' lives twenty-four hours a day. School-networked tablets, computers, and smart phones are used by students inside and outside of school.

So testing need not be the all-at-once, high-stakes affair it has always been. Old forms of testing are based on outdated needs to process pencil-and-bubble answer sheets. These practices: getting a large number of students in the same room at the same time to take the same test, which is then collected and processed—these could be supplemented or replaced by a more pervasive and constant technology-enabled monitoring. The future of testing will allow for ongoing evaluations of students' educational-opportunity use, engagement, and performance. The data collected will be unprecedented in its scope and will be housed permanently, transferable form school to school, and on up through higher education and into employment. This multi-metric

panoptical portfolio—this standardized educational biography—could come to replace the report card and diploma as a representation of educational attainment.

But surveillance and privacy are not the predominant concern for tomorrow's testing infrastructures. Besides, many of the issues raised are not distinctively educational ones, as violations of privacy can lead to violations of rights even more fundamental than the right to an education. The issues that concern us here involve the use of these new testing tools in the context of a new generation of students and teachers. How will the rapid proliferation of recent advances in psychology and neuroscience impact the design of new testing technologies? Moreover, what institutional policies and norms will be in place as testing technologies once again restructure schools?

The new sciences of learning and the future of testing

> It is easy for science to be regarded as a guarantee that goes with the sale of goods rather than as a light to the eyes and a lamp to the feet. It is prized for its prestige value rather than as an organ of personal illumination and liberation. It is prized because it is thought to give unquestionable authenticity and authority to a specific procedure to be carried out in the school room. So conceived, science is antagonistic to education as an art.
>
> -John Dewey (1929, p. 7)

Many who are opposed to testing and opposed to the current shape of educational reforms see these injustices as stemming from 'scientific' testing and so begin to question the very idea of a scientific approach to education at all. Critical theories of education often argue that the learning sciences have been a predominately negative force in schools (Apple, 2001). But there has, in fact, never been a true science of education brought into the schools. Like Christianity and Communism, as the saying goes, a true science of education has not failed—it has never been tried. Given the significance of advances in the learning sciences in recent decades, the possibilities for adopting them for use in test design are profound. Moreover, given the technology trends just discussed, it would seem that the time is ripe for seriously questioning the scientific and technological foundations of testing infrastructures.

Testing infrastructures depend on conceptions about the nature of learning. Often these theories of learning are not made explicit, either because they are under-theorized, or because there is, in fact, no theory of learning that guides design and implementation (NRC, 2011). This is deeply problematic because in order to use tests effectively and knowledgeably, those involved need to understand the meaning of the score a student receives. Unfortunately, this kind of a-theoretical approach to testing is more common than it might seem (*ibid*.). Recall the truncation of objectivity that characterizes the SAT and the related explanatory and pedagogical inarticulacies that stem from focusing merely on predictive validity.

Most test results provide an account of which problems a student got right and which ones they got wrong, although many (like the SAT) provide merely a single numeric score without an item-by-item breakdown of student performance. While it might seem that knowing which problems a student got right or wrong is valuable, in most cases this provides almost no insight into *why* a student answered the way they did or what could be done to improve student

learning. In the case of a single holistic numeric score, the results are even less informative. These limitations on the meaning of test scores arise when tests are not built in terms of an explicit theory of learning. That is, tests not designed in terms of a theory of learning simply cannot provide insight into *why* a student is unable to answer certain questions; they provide information only about which ones they got wrong (and sometimes they do not even provide that). So while an a-theoretical test can provide insight into the fact that a student struggles with certain kinds of math problems, about fractions for instance, it does not do so in terms of a system of ideas about how students learn to understand fractions. The test says, in effect, "This student is bad at fractions," but it gives no other insight into the learning processes that are in play (e.g., Is it because they never understood division properly, or because they switched the definitions of 'denominator' and 'numerator,' or what?).

Moreover, a test built without an explicit theory of learning can serve only very limited functions. Tests that are not wedded to an explicit theory of learning can be used to classify students and schools, to perform value-added analyses, and for a wide range of other efficiency-oriented functions, but these tests cannot be used as educative aids because they provide no insight into the learning process *per se*. On the other hand, building tests around an explicit theory of learning increases the range of functions a test can serve, allowing for insights that are directly beneficial and relevant to both student and teacher.

Unfortunately, the history of testing offered above shows that advances in learning theory have never driven the adoption and design of new testing technologies. Rather, concerns about efficiency and system-level accountability have historically trumped concerns about teaching and learning. During the century that gave us Dewey and Piaget (among many others!), testing infrastructures changed primarily in response to advances in technology enabling economies of

scale and the decision-making needs of bureaucrats—not in response to advances in the learning sciences that were progressively revealing the nature of how educational processes should be structured. That is, while advances in educational and developmental psychology were opening possibilities for approaches to assessments based on meaningful, student-centered, and psychologically realistic systems of categories and constructs, the dominant approach to testing remained focused on measuring under-theorized constructs (e.g., scholastic aptitude) using simplistic means (e.g., multiple-choice).

As discussed in Chapter 5, evidence continues to mount concerning the detrimental effects of these psychologically naïve testing practices, especially their stigmatizing and disempowering impact on students, and their tendency to radically truncate the pedagogical and curricular options available to teachers (RAND, 2011). It is likely that new educational technologies will be used primarily to deliver the same old multiple-choice tests, only faster, to more students, with greater frequency, and using more sophisticated data-analytic techniques. The short list of emerging edu-tech trends presented above, including new possibilities for multitudinous embedded, real-time assessment, could result in testing infrastructures that are increasingly insensitive to the needs of teachers and students, divorced from research about learning, and used mainly for ongoing systems-level surveillance of teacher, student, and school performance.

And yet these same technological trends open up possibilities for radically new testing approaches built around the best of what we know from the learning sciences. Building justice-oriented tests that can truly benefit students and teachers will involve explicitly adopting methods and theories from the learning sciences and infusing them into research and development efforts as well as test administration practices. This cuts against the grain of the

history of testing insofar as an explicit theory of learning would drive test design and technology adoption. It is now possible to harness new technologies to create a new kind of testing infrastructure that is learning- and learner-centric, thus serving social justice.

Theories of learning have become increasingly sophisticated in recent decades due to advances in cognitive science and neuroscience (Fischer, 2010). Likewise, advances in psychometrics have made it possible to reliably and validly measure a wider range of dynamic and meaningful constructs (Bond & Fox, 2001; Dawson, 2008). Contemporary testing infrastructures, and the uses to which they are put, reflect these advances in only a very limited way, if at all. The relations between the learning sciences and testing are long and complex, and a full treatment would require another book. The idea here is only to suggest the most basic ways that the learning sciences need to be brought into the design of justice-oriented testing infrastructures. This discussion is even more pressing given the tech-trends impacting testing outlined in above, which are setting the stage for a wholesale redesign of testing in the coming decades. Simply put, the future of testing hinges on questions about how to use the learning sciences to inform the design and adoption of new testing technologies (NRC, 2001; NSF, 2008).

To be clear: the argument here is *not* that a certain level of scientific knowledge is needed to build a justice-oriented testing infrastructure. Rather, the argument is that those seeking to design such a justice-oriented infrastructures are obligated to use best of what scientific knowledge is available about the nature of learning. Moreover, one need not even be committed to a particular theory of learning to argue for a testing infrastructure that is designed to make use of the learning sciences. The goal here is simply to outline some general parameters that could assure testing infrastructures be designed and continually revised in terms of advances in the learning sciences.

As the list of design principles below makes clear, aligning testing infrastructures with the learning sciences goes a long way toward securing the kind of testing technologies that would enable justice-oriented testing. This is not a coincidence. Throughout this work the main lesson about justice-oriented testing infrastructures—and the prime way they differ from efficiency-oriented ones—has been that they are relevant and beneficial to students and teachers. As just discussed, tests that are built around a theory of learning are much more likely to provide the necessary relevance and benefits. Of course even the best measures need to be institutionalized correctly to assure their just use, which is why the concluding section below addresses the governance and authority structures of schools. Nevertheless, no matter what the organization of authority in schools, standardized tests without a theory of learning behind them will almost always be useful only for efficiency-oriented purposes.

Testing infrastructures that combine the learning sciences with new trends in educational technologies would at least conform to the following design principles. This list is meant to be suggestive, not exhaustive; it is based on earlier work by Dawson & Stein (2011) where a fuller justification for their adoption can be found:

Evidence-based: Priorities should be shifted so that test development is guided by the learning sciences and informed by research about learning. This contributes to justice by providing students and teachers with results that reflect a psychologically realistic system of categories and thus more directly benefits teaching and learning.

Build knowledge: Given trends toward increasingly large digital databases of electronic learning records, tests should aim to contribute to the learning sciences through data

collection and housing. Justice is served insofar as advancing the sciences of learning improves teaching and learning.

Broadly available: Given trends toward technology saturation and online education, tests should leverage the Internet and computing technologies to serve the least-advantaged. Distributing advances in educational technology according to the difference principle is a social justice boon.

Support teaching and learning: Tests should enable customization, scaffolding, and just-in-time learning by organizing the delivery of educational resources to educators, students, and parents. This promotes the fair distribution of educational primary goods and directly benefits teachers and students.

Low-stakes: Just-in-time learning allows for multitudinous embedded assessments: many low-stakes formative tests on diverse topics with no high-stakes testing anxiety. Low-stakes testing is more fair to diverse student abilities, yields more objective data, and is less likely to result in school cultures burdened by the (psychological and financial) costs of intensive surveillance.

Relevant: Tests should leverage the diversity and ever-expanding affordances of online educational environments to allow students to operate on knowledge that matters to them and practice essential life skills while also working toward a mastery of the academic competencies targeted by standards. This allows testing to better provide for equal opportunity and self-actualization.

Embeddable: Tests that enable just-in-time learning and reflection should be made an integral part of classroom lessons, not tacked on at the end. Such tests benefit

teaching and learning, not merely organizational efficiency, bringing direct benefits to students and teachers.

Formative: Tests that enable customization and scaffolding should be learning experiences in themselves, directly contributing to student understanding. If the goal is to serve teachers and students first, then feedback from the test should be directly educative.

Diagnostic: Tests that enable customization, scaffolding, and reflection should be based on research into student learning, providing insights into what the student can do now and what would most benefit their future learning. Diagnostic information must be grounded in learning theory; it is directly relevant and beneficial to students and teachers.

Standardized: Increasingly large digital databases of electronic learning records will need to be built around common measures and indexes to enable both scientific and organizational learning; tests should thus be standardized to a universal learning scale. Objectivity and standardization are a necessary part of any just testing infrastructure.

These design parameters clearly require more elaboration than space provides for here, but they should give a rough sense of the new possibilities for testing that have been emerging at the interface of the learning sciences, burgeoning educational technologies, and social justice. The history of testing has taught that while new technologies bring transformations in testing practices, these transformations have occurred in isolation from the learning sciences. The

invention of the multiple-choice question, and then the Scantron machine, did more to shape testing than any advances in our understanding of learning. Students and teachers have paid the price while the demands of organizational efficiency and systems-level accountability have consistently trumped concerns about teaching and learning in the design and adoption of new technologies for testing. As the stage is now set for a new technology-wrought revolution of testing practices, it is critical that educators and policymakers build consensus around a set of design parameters like the ones above in order to shape the new testing infrastructure in ways that will be more beneficial to teachers and students.

Democracy, education, and the future of testing

> Democratic theory faces up to the fact of difference in our moral ideals of education by looking toward democratic deliberations not only as a means to reconciling those differences, but also as an important part of democratic education.... It makes a democratic virtue out of our inevitable disagreement over educational problems. The democratic virtue, simply stated, is that we can publicly debate educational problems in a way much more likely to increase our understanding of education and each other than if we were to leave the management to schools, as Kant suggests, "to depend entirely on the judgment of enlightened experts." The policies that result from our democratic deliberations will not always be the right ones, but they will be more enlightened than those that would be made by unaccountable educational experts.
>
> -Amy Gutmann (1999, p. 11)

The design principles introduced above suggest preferable directions for emerging technology-intensive testing infrastructures. But it is not enough to create new and better tools. The lessons learned in this work suggest that schools themselves need to be structured differently if they are to allow for the institutionalization of justice-oriented testing infrastructures. This final section explores these issues. The goal is to sketch the rough outlines of the school cultures and policies that would be needed to enable preferable futures for testing. Importantly, the suggestions offered below are worth adopting for reasons aside from their contribution to justice-oriented testing. But the account here focuses only on why these kinds of reforms are needed as a part of instituting a more just future for testing. Broader arguments in favor of democratizing schools are not discussed, but can be found in most of the citations offered. It will be clear, as it was in the section immediately above, that the account offered here is merely suggestive and evocative, not comprehensive and conclusive. More work needs to be done to justify both the design principles and the democratic forms of school governance that are the focus of this chapter.

There are at least two reasons why justice-oriented testing infrastructures require distinctly democratic forms of school organization and governance. The first concerns the creation of school cultures that are conducive to ongoing scientific research. Science is at its very core a radically democratic undertaking (Dewey 1929; Elgin, 1996). Schools that are undemocratic will be unscientific in so far as they systematically render mute the insights and perspectives of entire stakeholder groups (i.e., teachers and students). The second reason that justice-oriented testing requires a reconfiguration of authority in schools concerns the second and third principles of just institutional measurement—i.e., the need for measurement infrastructures

to be relevant and beneficial to everyone requires the institutionalization of means for determining their successes and failures in this regard. Without a way of communicating with those affected by them, the implementation of testing infrastructures will remain insensitive to those it most intimately involves, likely leading to the kinds of injustices discussed in previous chapters.

Comparing democratic decision making to scientific inquiry is a useful way to counteract the common sense idea that democracy is simply a matter of voting and that each vote is equally important. The form of democratic participation that is suggested here as an ideal for school governance is often referred to as *deliberative democracy* (Habermas, 1996; Gutmann, 1999). Like a scientific community, a deliberative democratic community is guided by "the unforced force of the better argument" and is committed to an un-coerced and free exchange of ideas in pursuit of solutions agreeable to all (Habermas, 1996). This kind of democratic decision-making does not entail that anything a majority happens to think will become policy or that everyone's opinion is somehow equally valid. Quite the opposite. Like scientific inquiry, deliberative democratic decision-making in schools "both empowers and constrains community control over education" (Gutmann, 1999 p.72). These forms of decision-making aim to promote reasoned debate in public deliberations about topics that are of general interest to the community. Already, deliberative democratic forms of school governance have been theorized, implemented, and studied (Apple & Beane, 1995; Buck & Villines, 2007).

However, this tradition in favor of democratizing schools has overlooked the argument that these forms of school organization are valuable on scientific as well as political grounds. In other words, if the goal is to make schools more scientific, to promote greater and more secure forms of objectivity through testing, and to thus generate usable knowledge about education

practice—then the schools should be structured like research communities, which are necessarily characterized by democratic methods of decision-making and deliberation. Of course, the ideal of a scientific community is as abstract as the ideal of a democratic one, and comparisons between the two do break down at various points (e.g., scientists need training before being officially counted as members of a research community). The sociology of science has shown since the 1950s that the working lives of scientists are far from the ideals of scientific practice that grace the pages of textbooks and the early work of philosophers of science. However, this same literature shows that science proceeds by insights that can come from anywhere, and that while the ideal of un-coerced, free deliberation is often far from realized, it nevertheless serves scientists as a reference point that is more than an ideological fiction. That is, while scientific practices may fall short of their ideal, it is nevertheless this ideal that guides practices and that is appealed to when critiquing breaches of practice (e.g., if it is discovered that a competent researcher's findings were not considered because they were unpopular, the community will see this as a breach of epistemic conduct).

The same cannot be said of most schools. In general, there are not generally accepted ideals of how authority and decision-making ought to be structured in schools. School governance structures range from radically authoritarian to radically egalitarian, and debates rage about what forms are preferable (Apple, 2001). The modest recommendation offered here is that these debates should include concerns about the scientific integrity of school cultures and the related possibilities for justice-oriented testing infrastructures. While discourse focuses on the interface of teacher unionization, educational de-professionalization, and impending school privatization, too few have their eyes on the impact of these trends on the science of education and the future of testing.

Imagine two schools that claim commitments to the science of education and to objective testing. One school is run as a deliberative democracy in which "participatory structures are put in place... [so that] teachers share in the 'management' or 'ownership' of the school" (Gutmann, 1999, p. 82). Teachers are brought into democratic deliberations about all school-wide policies while also given the power to inform research and test design in collaboration with experts (and in some cases through the acquisition of new expertise). In the other school, teachers are not included in decision-making about school-wide policies and are simply told about major changes, including the implementation of testing infrastructures and research being undertaken by experts with whom they have minimal contact. The question here (for the sake of argument) is not which school will do best in promoting learning, retaining teachers, or other important factors that bear on the success of a school's culture. The question is simply: *Which school would produce better scientific results?*

While this is an empirical question for which there is no answer, there is good reason to think it would be the democratically run school. For one, the results would likely be more trustworthy because there would be broad community support for and participation in the research as well a collective understanding of the conditions that secure good scientific results. The undemocratic school, on the other hand, faces all the problems of surveillance and enforcement discussed in Chapters 3 and 5, which lead to the decline of objectivity and the inefficiencies of injustice. The undemocratic school is likely to create inefficiencies in the production of usable scientific knowledge, both due to the decline of objectivity and the disconnect of research from teacher practice. Who knows better than teachers what educational questions deserve scientific attention and which answers are reasonable and likely to be adopted? It is for just these reasons that the need for teachers to be brought in as collaborators in scientific

investigations has been a resounding refrain in contemporary educational research (Hinton & Fischer, 2008; NRC, 2001).

There is also the question of which school is more likely to implement a justice-oriented testing infrastructure. Justice-oriented testing requires that tests be relevant and beneficial to everyone. The democratic school will have participatory structures that are sensitive and responsive to the needs of teachers, and would thus be capable of adjusting the properties of the testing infrastructure accordingly, both designing it in light of professed needs and revising it in light of experienced impacts. The undemocratic school would lack these dynamic methods for staying in touch with the needs and experiences of teachers. This is the second reason in favor democratizing schools discussed here: justice requires processes for taking account of the needs and experiences of those who are implicated in testing infrastructures.

Importantly, this requires that the participatory governance structure of schools be expanded to include students and parents. Expanding the participatory structures in this way is a complex topic, but models do exist (Apple & Beane, 1995; Buck & Villines, 2007). Clearly, there are a host of issues, ranging from the immaturity of students (e.g., they don't know what is good for them yet) to the self-interest and biases of parents (e.g., what most benefits their *own* children is taken as most important). Issues like these make things more complex, but they are not insurmountable barriers to this form of democratic participation (Gutmann, 1999). The point here is obviously not to settle this debate or to articulate a means for adjudicating the limits of extending democratic authority to students and teachers (although see: *ibid*.). Rather the goal is to simply point out that justice-oriented testing requires something along these lines. Parents, students, and teachers are those most affected by testing infrastructures, so they must have a role in shaping them. Exactly what this role ought to be requires a great deal more work, and likely a

great deal of experimentation and trial-and-error. This section has a much more modest task, which is only to point to the general direction reforms ought to proceed in if justice-oriented testing is to flourish in schools of the future.

In this work I have addressed foundational concerns at the interface of educational measurement and social justice. Following John Rawls's philosophical methods, I conceived and justified an ethical framework for guiding practices involving educational measurement. I demonstrated that educational measurement is critical in ensuring or inhibiting just educational arrangements. I also clarified a principled distinction between efficiency-oriented testing and justice-oriented testing. In order to test the feasibility and utility of my proposed framework, I employed it to analyze several historical case studies that exemplify the ethical issues related to testing: (1) the widespread use of IQ-style testing in US schools during the early decades of the twentieth century; (2) the founding of the Educational Testing Service; and (3) the recent history of test-based accountability associated with No Child Left Behind. I concluded with a set of speculative design principles and arguments in favor of radical democratic reforms, which address how the future of testing might be shaped to ensure justice for all. Perhaps it is a sign of a successful work that its conclusions demand that a great deal more work be done.

Bibliography:

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Alder, K. (2002). The measure of all things: the seven year odyssey that transformed the world. New York: The Free Press.

Apel, K. (1984). *Understanding and explanation: a transcendental-pragmatic perspective*. Cambridge: MIT Press.

Apple, M.W. (2001) *Educating the "right" way: markets, standards, God, and inequality*. New York: Routledge.

Apple, M.W. (2004). *Ideology and curriculum.* New York: Routledge

Apple, M.W. & Bean, J.A. (Eds) (1995). *Democratic schools.* Alexandra, VA: Association for Supervision and Curriculum Development.

Arrow, K., Bowles, S., & Durlauf, S. (Eds.) (2000). *Meritocracy and economic inequality.* Princeton: Princeton University Press.

Bartiz, L. (1965). *The servants of power: a history of the use of social science in American industry*. New York: John Wiley & Sons.

Bhaskar, R. (2013). *The possibility of naturalism: a philosophical critique of the contemporary human sciences.* New York: Routledge.

Binet, A. (1909/1973). *Les idees modernes sur les enfants* (with a preface by Jean Piaget). Paris: Flaammarion.

Black, H. (1964). *They shall not pass* New York: Morrow.

Block, N., & Dworkin, G. (Eds.). (1976). *The IQ controversy*. New York: Pantheon.

Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics*. New York: Cambridge University Press.

Bowles, S. & Gintis, H. (1976/2011). *Schooling in capitalist America: educational reform and the contradictions of economic life*. New York: Basic Books.

Bowles, S. & Gintis, H. (1986). *Democracy and capitalism: property, community, and the contradictions of modern social thought.* New York: Basic Books.

Bowles, S. & Gintis, H. (1998). *Recasting egalitarianism: new rules for communities, states, and markets.* New York: Verso.

Brown, J. (1992). *The definition of a profession: The Authority of a metaphor in the history of intelligence testing*. Princeton Princeton University Press.

Buchanan, A., Brock, D.W., Daniels, N., & Wikler, D. (2000). *From chance to choice: genetics and justice.* New York: Cambridge University Press.

Buck, J., & Villines, S. (2007). *We the People: Consenting to a Deeper Democracy; A Guide to Sociocratic Principles and Methods*. Washington DC: Sociocracy.info.

Busch, L. (2011). *Standards: recipes for reality*. Cambridge: MIT Press.

Brigham, C.C. (1923). *A study of American intelligence.* Princeton: Princeton University Press.

Brigham, C.C. (1937). The place of reseach in a testing organization. *School and society.* Vol. 46, no. 1198. pp. 756-759.

Bruner, J. (1960). *The process of education*. Cambridge: Harvard University Press.

Callahan, R. E. (1962). *Education and the cult of efficiency: a study of the social forces that have shaped the administration of the public schools*. Chicago: University of Chicago Press.

Campbell, D. T. (1975). Assessing the impact of planned social change. In G. M. Lyons (Ed.), *Social research and public policy: the Dartmouth/OECD Conference* Hanover, NH: Public Affairs Center, Dartmouth College.

Cappon, L. (Ed.). (1959). *The Adams-Jefferson Letters*. University of North Carolina Press.

Chapman, P. (1988). *Schools as sorters: Lewis M. Terman, applied psychology, and the intelligence testing Movement, 1890-1930*. New York: New York University Press.

Collins, A., & Halverson, R. (2009). *Rethinking education in the age of technology: the digital revolution and schooling in America*. New York: Teachers College Press.

Colman, A. M. (2001). *Dictionary of psychology*. New York: Oxford University Press.

Conant, J. B. (1970). My several lives: memoirs of a social inventor. New Yok: Harper & Row.

Conant, J.B. (1943). Wanted: American radicals. *The Atlantic monthly*. May, p. 41.

Connell, M., Stein, Z., & Gardner, H. (2012). Bridging between brain science and educational practice with design patterns. In Della Sala & Anderson (Eds.) *Neuroscience in education*. (pp. 267-286). Oxford University Press.

Cremin, L. (1969). *The transformation of the school*. New York: Knopf.

Cremin, L. (1970). *American education: the colonial experience, 1607-1783.* New York: Harper & Row.

Cremin, L. (1980). *American education: the national experience, 1783-1886.* New York: Harper & Row.

Cremin, L. (1988). *American education: the metropolitan experience, 1876-1980.* New York: Harper & Row.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30*, 1-14.

Danziger, K. (1990). *Constructing the subject: historical origins of psychological research*. New York: Cambridge University Press.

Darling-Hammond, L. (2010). *The flat world and education: how America's commitment to equity will determine our future.* New York: Teacher's College Press.

Dewey, J. (1916). *Democracy and education*. New York: The Macmillan Company.

Dewey, J. (1929). *The sources of a science of education*. New York: Liveright.

Dickson, V. E. (1922). Classification of school children according to mental ability. In Terman (Ed.). *Intelligence tests and school reorganization.* (pp. 32-52). New York: World book.

Duncan, A. (2011). Statement concerning Atlanta cheating scandal. Interview with 11Alive news, Atlanta. June 6th, 2011. Retrieved May, 2014, from: http://www.11alive.com/news/article/196896/40/Secretary-of-Education-stunned-by-scandal

Duncan, O. D. (1984). *Notes on social measurement: historical and critical*. New York: Russell Sage Foundation.

Dupree, A.H. (1957). *Science in the federal government*. Cambridge: Harvard University Press.

Durn, M.W. (1993). An A is not an A is not an A: a history of grading. *The educational forum*. *51*.

Elgin, C. (1996). *Considered Judgment*. Princeton, NJ: Princeton University Press.

Finn, C.E., & Petrilli, M.J. (2007). Introduction. In John Cronin et al. *The proficiency illusion*. Washington, D.C.: Fordham Institute and Northwest Evaluation Association.

Firestone, W. A., Frances, L., & Schorr, R. Y. (Eds.). (2004). *The ambiguity of teaching to the test: standards, assessment, and educational reform*. Mahwah, NJ: Erbaum Associates.

Fischer, K. W. and Bidell, T.R. (2006). Dynamic development of action, thought, and emotion. In W. Damon and R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical*

models of human development (pp. 313-399). New York, Wiley.

Forge, J. (Ed.). (1987). *Measurement, realism, and objectivity*. Boston: D. Reidel Publishing.

Freedman, S. (2007). *Rawls*. New York: Routledge.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning *American Psychologist, 39*, 193-202.

Gardner, H. (2011). *Frames of mind: the theory of multiple intelligences*. New York: Basic Books.

Georgia Bureau of Investigation (2011). *Special Investigation into test tampering and related matters in the Atlanta Public Schools*. Atlanta: Office of the Governor.

Gifford, B. R. (Ed.). (1989). *Test policy and the politics of opportunity allocation: the workplace and the law*. Boston: Kluwer Academic.

Goddard, H.H. (1913). The Binet test in relation to immigration. *Journal of psycho-asthenics.* 18: pp. 105-107.

Goddard, H.H. (1914). *Feeble-mindedness: its causes and consequences.* New York: Macmilian.

Goddard, H.H. (1920). *Human Efficiency and levels of intelligence.* Princeton: Princeton University Press.

Gould, S. J. (1996). *The mismeasure of man (revised and expanded edition)*. New York: Norton.

Gross, M. (1962). *The brain watchers*. New York: Random House.

Gutmann, A. (1984). *Democratic education.* Princeton, NJ: Princeton University Press.

Habermas, J. (1971). *Knowledge and human interests*. Boston: Beacon Press.

Habermas, J. (1984). *The theory of communicative action:reason and the rationalization of society.* (T. McCarthy, Trans. Vol. 1). Boston: Beacon Press.

Habermas, J. (1987). *The theory of communicative action: Lifeworld and system, a critique of functionalist reason* (T. McCarthy, Trans. Vol. 2). Boston: Beacon Press.

Habermas, J. (1988). *On the logic of the social sciences* (Nicholsen & Stark, Trans.). Cambridge: MIT Press.

Habermas, J. (1990). *Moral consciousness and communicative action*. (Nicholsen & Stark, Trans.). Cambridge: MIT Press.

Habermas, J. (1996). *Between facts and norms: contributions to a discourse theory of law and democracy* (W. Rehg, Trans.). Cambridge, MA: MIT Press.

Haney, W. M. (1981). Validity, vaudeville, and values: a short history of social concerns over standardized testing. *American Psychologist, 36*, 997-1000.

Haney, W. M., Madaus, G. F., & Lynos, R. (1993). *The fractured market place for standardized testing*. Norwell, MA: Kluwer Academic.

Hayden, P. (2002). *John Rawls: toward a just world order.* Cardiff: University of Wales Press.

Herrnstein, R.J., & Murray, C. (1994). *The bell curve: the reshaping of American life by difference in intelligence*. New York: Free Press.

Hess, F, & Finn, C.E. (Eds.) (2007). *No remedy left behind: lessons from a half-decade of NCLB.* Washington, D.C.: AEI Press.

Hess, F., & Petrilli, M. (2006). *No Child Left Behind*. New York: Peter Lang.

Hinton, C. and Fischer, K. W. (2008), Research Schools: Grounding Research in Educational Practice. *Mind, Brain, and Education*, 2: 157–160.

Houts, P. L. (1977). *The myth of measurability*. New York: Hart Publishing.

Hunter, J. S. (1980). The national system of scientific measurement. *Science, 210*(4472), 869-874.

Hursh, D. (2008). *High-stakes testing and the decline of teaching and learning*. New York: Rowman & Littlefeild.

Jaques, E. (1976). *A General Theory of Bureaucracy*. London: Heinemann Educational.

Johnson, D. D. (2008). *Stop high-stakes testing: an appeal to America's conscience*. Lanham: Rowman & Littlefeild Publishers.

Kant, I. (1785/1996). *Groundwork of the metaphysics of morals*. New York: Cambridge Univeristy Press.

Kevles, D. (1998). *In the name of eugenics: genetics and the uses of human heredity. Cambridge*: Harvard University Press.

Kohlberg, L. and D. Candee (1984). Stage and sequence: The cognitive developmental approach to socialization. *Essays on moral development: Vol. 2. The psychology of moral development: The nature and validity of moral stages* (pp. 7-169). San Francisco: Jossey Bass.

Koretz, D. M. (2008). *Measuring up: what educational testing really tells us*. Cambridge: Harvard University Press.

Kozol, J. (2012). Savage inequalities. New York: Harper & Row.

Kuhn, T. (1962). The structure of scientific revolutions. Chicago: University of Chicago Press.

Kula, W. (1986). *Measures and men*. Princeton: Princeton University Press.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science, 315*, 1080-1081.

Lagemann, E. (2000). *An elusive science: the troubling history of educational research*. Chicago: University of Chicago Press.

Lane, R. (1991). *The market experience.* New York: Cambridge University Press.

Lawler, J. M. (1978). *IQ, heritability, and racism*. New York: International publishers.

Lazarus, M. (1981). *Goodbye to excellence: a critical look at minimum competency testing*. Boulder, CO: Westview Press.

Lemann, N. (1999). *The big test: the secret history of the American meritocracy*. New York: Farrar, Straus and Grioux.

Liebenau, J. (1987). *Medical science medical industry*. London: MacMillan Press.

Lippmann, W. (1922). The Lippmann-Terman debate. In Block & Dworkin (Eds.) *The IQ controversy.* (pp. 4-44). New York: Pantheon Books.

Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing.

Luce, R. D. (1972). What sort of measurement is psychophysical measurement. *American Psychologist, 26*(2), 96-106.

Maslow, A. (1968) *Toward a psychology of being.* New York: Van Nostrand Reinhold.

McMurrer, J. (2007). *Choices, changes, and challenges: curriculum and instruction in the NCLB era.* Washington, D.C: Center on Educational Policy.

McMurrer, J. (2008). *Instructional time in elementary schools: a closer look at changes for specific subjects.* Washington, D.C: Center on Educational Policy.

Messick, S. (1975). The standard problem: meaning and values in measurement and education. *American Psychologist, 30*, 5-11.

Messick, S. (1980) Test validity and the ethics of assessment. *American psychologist.* 35(11), 1012-1027.

Michell, J. (1999). *Measurement in psychology: A Critical history of a methodological Concept*. New York: Cambridge University Press.

Mills, C.W. (1956/1976). The sociological imagination. New York: Oxford University Press.

Minton, H. (1990). Lewis M. Terman and mental testing: in search of a democratic ideal. In M. Sokal (Ed.), *Psychological testing in American society: 1890-1930* (pp. 95-113). New Brunswick: Rutgers University Press.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory of a new generation of tests*. Hillsdale, NJ.: Erlbaum.

Montefinise, A. (2007). Lessons lost in test-prep craze. *New York Post*. January 28.

Mundy, B. (1987). The metaphysics of quantity. *Philosophical Studies, 51*, 29-54.

Nader, R. (1979). *Introduction.* In Nairn (author) *The rein of ETS: the corporation that makes up minds: The Ralph Nader Report on the Educational Testing Service*

Nairn, A. (1979). *The rein of ETS: the corporation that makes up minds: The Ralph Nader Report on the Educational Testing Service.*

National Research Council (1999). *High Stakes: testing for tracking, promotion, and graduation*. Washington, D.C.: National Academy Press.

National Research Council (2001). *Knowing what students know: the science and design of educational assessment*. Wahington, D.C.: National Academy Press.

NSF task force on cyberlearning. (2008). Fostering learning in the networked world: the cyberlearning opportunity and challenge. Washington, DC: National Science Foundation.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: how high-stakes testing corrupts America's schools*. Cambridge: Harvard Education Press.

Nobel, D. F. (1977). *America by design: science, technology, and the rise of corporate capitalism*. New York: Knopf.

Nussbaum, M. (2006). *Frontiers of justice: disability, nationality, species membership*. Cambridge, MA: Harvard University Press.

Obama, B (2008) Speech to the 146th Annual Meeting and 87th Representative Assembly of the

National Educational Association. Delivered July 5th, 2008.

O'Donnell, J. M. (1985). *The origins of behaviorism: American psychology 1870-1920*. New York: New York University Press.

Packard, V. (1964). *The naked society*. New York: McKay.

Phelps, R. P. (2003). *Kill the messenger: the war on standardized testing*. New Brunswick, N.J.: Transaction Publishers.

Piaget, J. (1932). *The Moral judgment of the child.* New York: Free Press.

Piaget, J. (1970). *The place of the sciences of man in the system of sciences*. New York: Harper & Row.

Phillips, K. R. (2003). *Testing controversy: a rhetoric of education reform* Cresskill, NJ: Hampton Press.

Porter, T. M. (1996). *Trust in numbers: the pursuit of objectivity in science and public life*. Princeton Princeton University Press.

Pariser, E. (2011). *The filter bubble: how the new personalized web is changing what we read and how we think.* New York: Penguin.

RAND (2010). *Reauthorizing No Child Left Behind: facts and recommendations*. Arlington, VA: RAND.

Ravitch, D. (2010). *The life and death of the great American school system: how testing and choice are undermining education*. New York: Basic Books.

Ravitch, D. (2013). *Reign of error: the hoax of the privatization movement and the danger to America's public schools.* New York: Knopf.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Rawls, J. (1996). *Political Liberalism*. New York: Columbia University Press.

Rawls, J. (1999). *Collected papers.* Cambridge, MA: Harvard University Press.

Rawls, J. (2000). *Lectures on the history of moral philosophy*. Cambridge, MA: Harvard University Press.

Rawls, J. (2001). *Justice as fairness*. Cambridge, MA: Harvard University Press.

Reed, J. (1990). Robert Yerkes and the mental testing movement. In M. Sokal (Ed.), *Psychological testing in American society: 1890-1930* (pp. 75-95). New Brunswick:

Rutgers University Press.

Resnick, D. P. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences and controversies* (pp. 173-194). Washington, D.C.: National Research Council.

Rhoades, K., & Madaus, G. (2003). Errors in standardized tests: a systemic problem. Chestnut Hill, MA: National Board on Educational Testing and Public Policy, Lynch School of Education, Boston University.

Ross, W.D. (Ed.) (1921). *The works of Aristotle.* Vol 10. Oxford: Clarendon Press.

Ryan, K. E., & Shepard, L. A. (Eds.). (2008). *The future of test-based educational accountability*. Newy York: Routledge.

Searle, J. (1995). *The construction of social reality.* New York: The Free Press.

Sokal, M. (Ed.). (1990). *Psychological testing in American society: 1890-1930*. New Brunswick: Rutgers University Press.

Spring, J. H. (1989). *The sorting machine revisited: national educational policy since 1945*. New York: Longman.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*(4), 215-227.

Sacks, P. (1999). *Standardized minds: the high price of America's testing culture and what we can do to change it*. Cambridge, MA: Perseus Press.

Samelson, F. (1990). Was early metal testing: a) racist inspired, b) objective science, c) a technology for democracy, d) the origin of the multiple choice exam, e) none of the above. In M. Sokal (Ed.), *Psychological testing in American society: 1890-1930* (pp.113-128). New Brunswick: Rutgers University Press.

Scanlon, T. M. (1998). *What we owe to each other*. Cambridge, MA: Harvard University Press.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven: Yale University Press.

Scheffler, I. (1960). *The language of education*. Springfield, Ill: Thomas.

Sen, A. (2000). Merit and justice. In Arrow, K., Bowles, S., & Durlauf, S. (Eds.) *Meritocracy and economic inequality.* Princeton: Princeton University Press.

Simon, B. (1971). *Intelligence, psychology and education*. London: Lawrence and Wishart.

Singh, I., & Rose, N. (2009). Biomarkers in psychiatry. *Nature, 460*(9), 202-207.

Smarter Balance Assessment Consortium (2014). Quarterly reports, April 2011-Dec 2013. Retrieved May, 2014, from: http://www.smarterbalanced.org/about/

Smith, L. (2006). Norms and normative facts in human development. In L. Smith & J. J. Voneche (Eds.), *Norms in human development*. New York: Cambridge University Press.

Stein, Z. (2007). Modeling the demands of interdisciplinarity: toward a framework for evaluating interdisciplinary endeavors. *Integral Review*, 4, 92-107

Strenio, A. (1981). *The testing trap*. New York: Rawson, Wade Publishers.

Swoyer, C. (1987). The metaphysics of measurement. In J. Forge (Ed.), *Measurement, realism, and objectivity: essays on measurement in the social and physical sciences* (pp. 235-290). Boston: D. Reidel Publishing.

Tavernor, R. (2007). *Smoot's ear: the measure of humanity.* New Haven, CT: Yale University Press.

Taylor, F. (1911). *The principles of scientific management.* New York: Harper & Row.

Thorndike, E.L. (1922). Measurement in education. In Whipple (ed). *The twenty-first yearbook of the National Society for the Study of Education: intelligence tests and their use.* Bloomington, IL: Public School Publishing Co. p. 1.

Toch, T. (2006). Margins of error: the education testing industry in the No Child Left Behind Era. Washington D.C.: Education Sector.

Tyack, D. B. (1974). *The one best system: a history of American urban education*. Cambridge: Harvard University Press.

US department of education. (2011). Race to the top assessment program Retrieved July, 2011, from http://www2.ed.gov/programs/racetothetop-assessment/index.html

Von Wright, G.H. (1971). *Explanation and understanding*. Ithaca, NY: Cornell University Press.

Wigdor, A. K., & Garner, W. R. (Eds.). (1982). *Ability testing: Uses, consequences and controversies*. Washington, D.C.: National Research Council.

Yerkes, R.M. (Ed.) (1921) Psychological examining in the United States Army. *Memoirs of the National Academy of Sciences,* vol. 15.

Zenderland, L. (1998). *Measuring mind: Henry Goddard and the origins of American intelligence testing.* New York: Cambridge University Press.

VITA

Zachary Stein

Degrees:

Jan 2004	*Hampshire College*, B.A., Philosophy / Cognitive Science
June 2006	*Harvard University Graduate School of Education*, Ed.M., Mind, Brain, and Education

Professional experience:

2002-2005	Research Assistant, *Collaboration for Excellence in Science Education (CESE)*, Hampshire College
2003-2006	Research Associate, *Developmental Testing Service* (formerly Sequence Consulting)
2006-2007	Teaching Fellow, Harvard University
2006-2010	Senior Analyst, *Developmental Testing Service*
2009-2010	Adjunct Faculty member, Integral Theory Department, John F. Kennedy University.
2010-2011	Deputy Director, *Developmental Testing Service*
2011-2014	Co-founder, Senior Outreach Liaison, Philosopher of Education, Lectica, Inc. (formerly Developmental Testing Service).
2008-2014	Senior Teaching Fellow, Harvard University
2014-present	Core Faculty and Associate Director of Assessment and Student Development, Meridian University
2014-presnet	Academic Director, Center for Integral Wisdom