Bend-La Pine Schools
School Board Work Session Meeting Agenda

September 27, 2022, 5:30 PM

Location:
Education Center, Board Room #314
520 NW Wall Street
Bend, OR 97703

1. Call to Order
   Speaker(s): Chair Barnes Dholakia
2. Pledge of Allegiance
   Speaker(s): Chair Barnes Dholakia
3. Review of Agenda 2
   Speaker(s): Chair Barnes Dholakia
   Description: Any changes to the Agenda after posting on September 23, 2022, are shown below.
   Attachments:

9.27.22 - Agenda de la Sesio´n de Trabajo - BORRADOR 2

4. Work Session 3
   Description: The Board will focus on key Board work and initiatives.
   Attachments:

Pre-Reading- The Tyranny of Metrics Ch 1-2 and 15-16 3

   A. System Performance Measures 17
      Speaker(s): Dave VanLoo, Director of School Improvement
      Attachments:

Presentation: Board Ends September 2022 17
Presentación: Junta Finaliza Presentación Septiembre 2022 23
DRAFT Board Ends 29
Inferences Activity 31

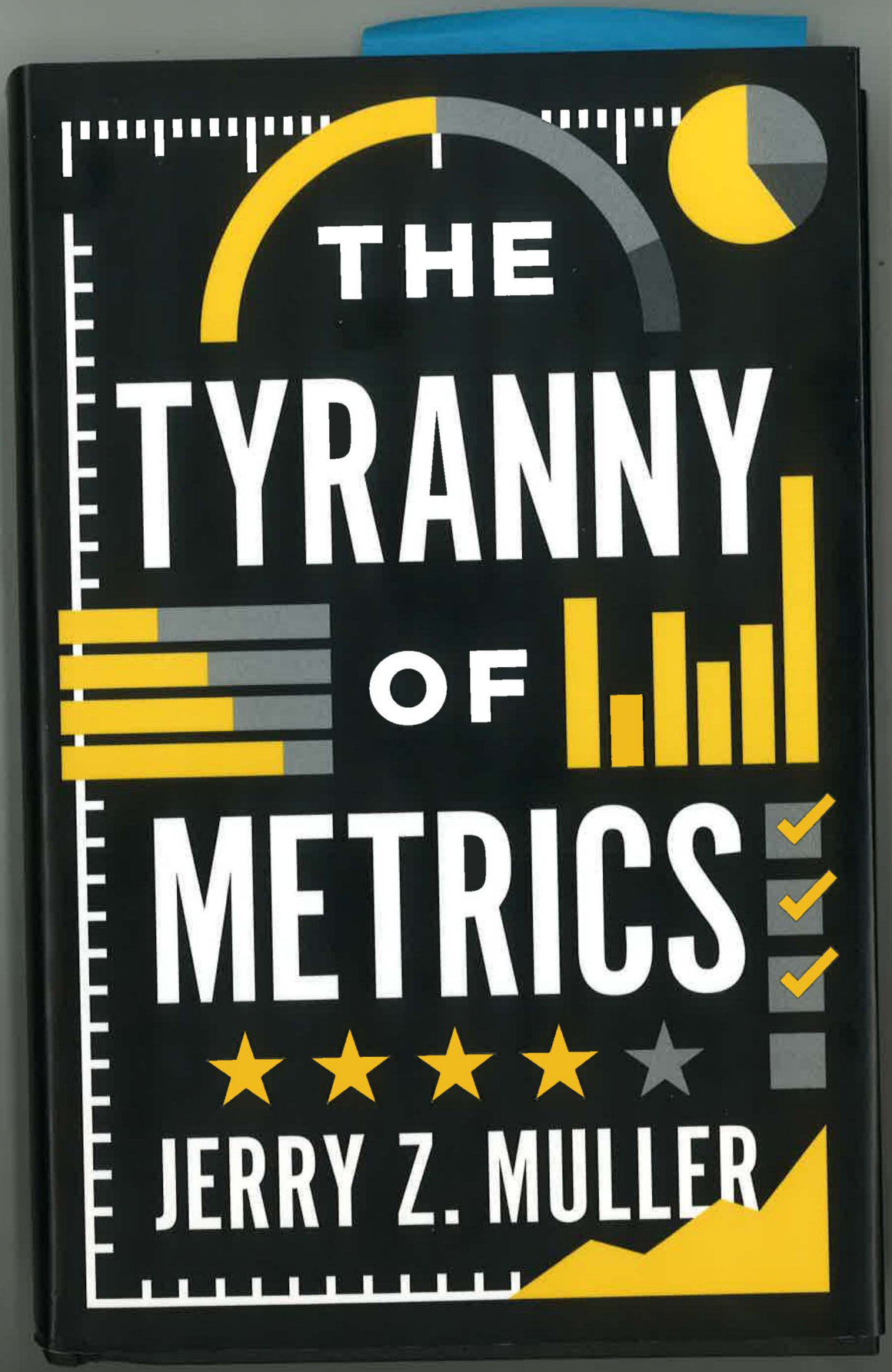
5. Board Comments
   Description: An opportunity for board members to provide comments or reflections.
6. Adjourn
   Description: Meeting will be adjourned with next Regular School Board Business Meeting scheduled for October 11, 2022.

Escuelas de Bend-La Pine
Agenda de la Sesión de Trabajo de la junta de la Mesa Directiva Escolar

27 de septiembre de 2022, 5:30 PM

Ubicación:
Centro de Educación, Sala de juntas #314
520 NW Wall Street
Bend, OR 97703

1. **Llamado al Orden**
   Ponente(s): Presidenta de la mesa directiva, Barnes Dholakia
2. **Juramento a la Bandera**
   Ponente(s): Presidenta de la mesa directiva, Barnes Dholakia
3. **Revisión de la Agenda**
   Ponente(s): Presidenta de la mesa directiva, Barnes Dholakia
   Descripción: Cualquier cambio en la Agenda después de la publicación el 23 de septiembre de 2022 se muestra a continuación.
4. **Sesión de Trabajo**
   Descripción: La mesa directiva se centrará en el trabajo y las iniciativas clave de la mesa directiva.
   A. Medidas de desempeño del sistema
      Ponente(s): Dave VanLoo, director de mejoramiento escolar.
5. **Comentarios de la mesa directiva**
   Descripción: Una oportunidad para que los miembros de la mesa directiva proporcionen comentarios o reflexiones.
6. **Levantamiento de la sesión**
   Descripción: La reunión se aplaza con la próxima Reunión Regular de Asuntos de la Mesa Directiva Escolar programada para el 11 de octubre de 2022.


THE
TYRANNY
OF
METRICS
JERRY Z. MULLER

1

THE ARGUMENT IN A NUTSHELL

There is a cultural pattern that has become ubiquitous in recent decades, engulfing an ever-widening range of institutions. Depending on taste, one could call it a cultural "meme;' an "episteme;' a "discourse;' a "paradigm;' a "self-reinforcing rhetorical system;'[1] or simply a fashion. It comes with its own vocabulary and master terms. It affects the way in which people *talk* about the world, and thus how they *think* about the world and how they *act* in it.[2] For convenience, let's call it metric fixation.

A key premise of metric fixation concerns the relationship between measurement and improvement. There is a dictum (wrongly) attributed to the great nineteenth-century physicist Lord Kelvin: "If you cannot measure it, you cannot improve it?' (n 1986 the American management guru, Tom Peters, embraced the motto, "What gets measured gets done;' which became a cornerstone belief of metrics.[3] In time, some drew the conclusion that "anything that can be measured can be improved:'[4]

When proponents of metrics advocate "accountability,' they tacitly combine two meanings of the word. On the one hand, to be accountable means to be responsible. But it can also mean "capable of being counted:' Advocates of "accountability" typically assume that only by counting can institutions be cruly responsible. Performance is therefore equated with what can be reduced to standardized measurements.

When proponents of metrics demand "transparency" they often insinuate that probity requires making explicit and visible as much information as possible. The result is the demand for ever more documentation, ever more mission statements, ever more "goal-setting [5]

The key components of metric fixation are

- the belief that it is possible and desirable to replace judgment, acquired by personal experience and talent, with numerical indicators of comparative performance based upon standatdized data (mettics);
- the belief that making such metrics public (transparent) assures that institutions are actually carrying out their purposes (accountability);
- the belief that the best way to motivate people within these organizations is by attaching rewards and penalties to their measured performance, rewards that are either monetary (pay-for-performance) or repl!ltational (rankings).

Metric fixation is the persistence of these beliefs despite their unintended negative consequences when they are put into practice.[6] It occurs because not everything that is important is measureable, and much that is measurable is unimportant. (Or, in the words of a familiar dictum, "Not everything that can be counted counts, and not everything that counts can be counted [7]) Most organizations have multiple purposes, and that which is measured and rewarded tends to become the focus of attention, at the expense of other essential goals. Similarly, many jobs have multiple facets, and measuring only a few aspects creates incentives to neglect the rest.[8] When organizations committed to metrics wake up to this fact, they typically add more performance measures-which creates a

cascade of data, data that becomes ever less useful, while gathering it sucks up more and more time and resources.

In the process, the nature of work is transformed in ways that are often pernicious. Professionals tend to resent the impositions of goals that may conflict with their vocational ethos and judgment, and thus morale is lowered. Almost inevitably, many people become adept at manipulating performance indicators through a variety of methods, many of which are ultimately dysfunctional for their organizations. They fudge the data or deal only with cases that will improve performance indicators. They fail to report negative instances. In extreme cases, they fabricate the evidence.

A frequent feature of metric fixation is paying for performance, that is, offering individuals or organizations financial incentives to meet quantifiable criteria. That may work in organizations that exist for the single purpose of making a profit, though as we'll see, even in these cases it is rarely effective. It works even less well in organizations in which employees are oriented to a more idealistic mission, such as schools, universities, medical practices, and hospitals. Whenever reward is tied to measured performance, metric fixation invites gaming.

Because the theory of motivation behind pay for measured performance is stunted, results are often at odds with expectations. The typical pattern of dysfunction was formulated in 1975 by two social scientists operating on opposite sides of the Atlantic, in what appears to have been a case of independent discovery. What has come to be called "Campbelli's Law,' named for the American social psychologist Donald T. Campbell, holds that "[t]he more any quantitative social 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.'[9]

In a variation named for the British economist who formulated it, we have Goodhart's Law, which states, "Any measure used for control is unreliable?'[10] To put it another way, anything that can be measured and Fewarded will be gamed. We will see many variations on this theme.

Trying to-force people to conform their work to preestablished_nu,mei;ieal goals tendno stifle iinnovation and creativity-valuable qualities in most settings. And it almost inevitably leads to a valuation of short-term goals over long-telim ptuposes.

In situations where there are no real feasible solutions to a problem, the gathering and publication of performance data serves as a form of virtue signaling. There is no real progress to show, but the effort demonstrated in gathering and publicizing the data satisfies a sense of moral earnestness. In lieu of real progress, the progress of measurement becomes a simulacrum of success. We'll see that in the case of the edueational "achievement gap?'

Because belief in its efficacy seems to outlast evidence that it frequently doesn't work, metric fixation has elements of a cult. Studies that demonstrate its lack of effectiveness are either ignored, or met with the assertion that what is needed is more data and better measurement. Metric fixation, which aspires to imitate science, too often resembles faith.

All of that is not intended to claim that measurement is useless or intrinsically pernicious. One of the purposes of this book is to specify when perfolimance metr.ics are genuinely useful-how to use metrics without the characteristic dysfunctions of metric fixation.

The next chapter, "Recurring Flaws;' provides a taxonomy of the most frequent types of flaws in the use of performance

metrics. Defining and labeling them will make it easier to refer back to them later. Then, in part II, we examine the origins of metric fixation and account for its spread and tenacity in spite of its frequent failures, in addition to exploring some of the deeper philosophical sources of its shortcomings. Part III comprises case studies that examine the more recent record of metrics, it successes and its shortcomings in a variety of fields, including K-12 education, higher education, medicine, policing, the military, business, and philanthropy and foreign aid. These case studies are intended to be suggestive rather than definitive. That is, they don't deal with every way in which the metric fixation manifests itself in each domain. Rather they provide concrete examples of recurring flaws and unintended consequences, as well as examples of the successful use of metrics from which we may derive lessons that can be applied in other domains. This section is followed by a brief excursus on the theme of transparency as the enemy of performance in certain realms. Finally, Part IV draws upon the preceding analysis to enumerate the unintended negative consequences of metric fixation and offer some guidelines about when and how to make use of metrics without succumbing to metric fixation.

2

RECURRING FLAWS

The drive to institute metrics often arises from the best of intentions, as a purported solution to real problems. And in some cases, as we'll see, it really *does* fulfill its promise to provide such solutions, or at least contributes to solving problems. But after decades of experience with the negative effects of metrics, as metric dysfunction threatens to cascade into yet more institutions, we should be able to anticipate the recurrent flaws. Here's a list to help identify and remember them. Of course, while we may distinguish them for purposes of analysis, these flaws often overlap in the real world.

Let's begin with problems of the *distortion* of information.

Measuring the most easily measurable. There is a natural human tendency to try to simplify problems by focusing on the most easily measureable elements.1 But what is most easily measuredr'is rarely what is most important, indeed sometimes not important at all. That is the first source of metric dysfunction.

Closely related is **measuring the simple when the desired outcome is complex.** Most jobs have multiple responsibilities and most organizations have multiple goals. Focusing measurement on just one responsibility or goal often leads to deceptive results.

Measuring inputs rather than outcomes. It is often easier to measure the amount spent or the resources injected into

a project than the *results* of the efforts. So organizations measure what they've spent, rather than what they produce, or they measure process rather than product.

Degrading information quality through standardization. Quantification is seductive, because it organizes and simplifies knowledge. It offers numerical information that allows for easy comparison among people and institutions.[2] But that simplification may lead to distortion, since making things comparable often means that they are stripped of their context, history, and meaning.[3] The result is that the information appears more certain, and authoritative than is actually the case: the caveats, the ambiguities, and uncertainties are peeled away, and nothing does more to create the appearance of certain knowledge than expressing it in numerical form.[4]

Campbell's Law and Goodhart's Law are warnings about the inevitable attempts to game the metric when much is at stake. Gaming the metrics takes a variety of forms.

Gaming through creaming. This takes place when practitioners find simpler targets or prefer clients with less challenging circumstances, making it easier to reach the metric goal, but excluding cases where success is more difficult to achieve.

Improving numbers by lowering standards. One way of improving metric scores is by lowering the criteria for scoring. Thus, for example, graduation rates of high schools and colleges can be increased by lowering the standards for passing. Or airlines improve their on-time performance by increasing the scheduled flying time of their flights.

Improving numbers through omission or distortion of data. This strategy involves leaving out inconvenient instances, or classifying cases in a way that makes them disappear from the metrics. Police forces can "reduce" crime rates by booking felonies as misdemeanors, or by deciding not to book reported crimes at all.

Cheating. One step beyond gaming the metrics is cheating—a phenomenon whose frequency tends to increase directly with the stakes of the metric in question. As we'll see, as the No Child Left Behind Act raised the stakes for schools of the test scores of their pupils, teachers and principals in many cities responded by altering students' answers on the test.

15

UNINTENDED BUT PREDICTABLE NEGATIVE CONSEQUENCES

One conception of the purpose of social science was articulated in the nineteenth century by Auguste Comte: *Savoir pour prevoir, prevoir pour prevenir* (Know in order to predict, predict in order to avert [the previously unanticipated consequences of our actions]). Now that we know a good deal about metric fixation, we can anticipate many of its unintended negative consequences, and perhaps avert them. Before we turn to the proper use of measured performance, let us gather together some lessons from our case studies about the recurrent perils of metrics.

Goal displaeemenUhrough 'liiversEon o feffort to *whagget"S' measu,red.* Goal displacement comes in many varieties. When performance is judged by a few measures, and the stakes are high (keeping one's job, getting a raise, raising the stock price at the time that stock options are vested), people will focus on satisfying those measures-often at the expense of other, more important organizational goals that are not measured.[1] Economists Bengt Holmstrom and Paul Milgrom have described it in more formal terms as a problem of misaligned incentives: workers who are rewarded for the accomplishment of measurable tasks reduce the effort devoted to other tasks.[2] The result

is that the metric *means* comes to replace the organizational *ends* that those means ought to serve.

Promoting short-termism. Measured performance encourages what Robert K. Merton called "the imperious immediacy of interests ... where the actor's paramount concern with the foreseen immediate consequences excludes consideration of further or other consequences?'[3] In short, advancing short-term goals at the expense of long-range considerations.

Costs in employee time. To the debit side of the ledger must also be added the transactional costs of metrics: the expenditure of employee time by those tasked with compiling and processing the metrics-not to speak of the time required to actually read them. That is exacerbated by the "reporting imperative"-the perceived need to constantly generate information, even when nothing significant is going on. Sometimes the metric of success is the number and size of the reports generated, as if nothing is accomplished unless it is extensively documented. Those within the organization end up spending more and more time compiling data, writing reports, and attending meetings at which the data and reports are coordinated. So, as the heterodox management consultants Yves Morieux and Peter Tollman note, employees work longer and harder at activities that add little to the real productiveness of their organization, while sapping their enthusiasm.[4]

Diminishing utility. Sometimes, newly introduced performance metrics will have immediate benefits in discovering poorly performing outliers.[5] Having gleaned the low-hanging fruit, there is tendency to expect a continuingly bountiful harvest. The problem is that the metrics continue to get collected from everyone. And soon the marginal costs of assembling and analyzing the metrics exceed the marginal benefits.

Rule cascades. In an attempt to staunch the flow of faulty metrics through gaming, cheating, and goal diversion, organizations institute a cascade of rules. Complying with them further slows down the institution's functioning and diminishes its efficiency.

Rewarding luck. Measuring outcomes when the people involved have little control over the results is tantamount to rewarding luck. It means that people are rewarded or penalized for outcomes that are actually independent of their efforts. Those penalized rightly feel that they've been treated unfairly.

Discouraging risk taking. Attempts to measure productivity through performance metrics have other, more subtle effects: they not only promote short-termism, as noted earlier, but also discourage initiative and risk-taking. The intelligence analysts who ultimately located Bin Laden worked on the problem for years. If measured at any point, their productivity would have seemed to be zero. Month after month, their failure rate was 100 percent, until they achieved success. From the perspective of their superiors, allowing the analysts to work on the project for years involved a high degree of risk: the investment in time might not have panned out. Yet really great achievements often depend on such risks. This is typical of situations involving long-term investments of manpower.

Discouraging innovation. When people are judged by performance metrics, they are incentivized to do what the metrics measure, and what the metrics measure will be some established goal. But that impedes innovation, which means doing something that is not yet established, indeed hasn't been tried out. Innovation involves experimentation. Trying out something new entails risk, including the possibility, perhaps prob-

ability, of failure.[6] When performance metrics discourage risk they inadvertently promote stagnation.

Discoura-gi'ng cooperation and common,.f!!!:..._rpose. Rewarding individuals for measured performance diminishes the sense of common purpose as well as the social relationships that provide the unmeasureable motivation for cooperation and institutional effectiveness.7 Reward based on measured performance tends to promote not cooperation but competition. If the individuals or units respond to the incentives created, rather than aiding, assisting, and advising one another, they strive to maximize their own metrics, ignoring, or even sabotaging, their fellows. As Donald Berwick, a leading medical reformer, has recounted,

> One hospital CEO described to me his system of profit-center management, in which middle management bonuses depended on local budget performance. I asked him if one of his managers would transfer resources from his department to another's if it would help the organization as a whole. "Yes; the CEO answered honestly, "if he were crazy.'[8]

Degradation of work. Compelling the people in an organization to focus their efforts on the narrow range of what gets measured leads to a degradation of the experience of work. Edmund Phelps, a Nobel Prize winning economist, claims in his book ***Mass Flourishing: How Grassroots Innovation Created Jobs, Challenge, and Change*** that one of the virtues of capitalism is its ability to provide "the experience of mental stimulation, the challenge of new problems to solve, the chance to try the new, and the excitement of venturing into the unknown:'[9] That is indeed a possibility under capitalism. But those subject to performance metrics are forced to focus their efforts on

limited goals, imposed by others, who may not understand the work that they do. For the workers under scrutiny, mental stimulation is dulled, they decide neither the problems to be solved nor how to solve them, and there is no excitement of venturing into the unknown because the unknown is beyond the measureable. In short, the entrepreneurial element of human nature-which extends beyond the owners of enterprises-may be stifled by metric fixation.[10]

One result is to motivate those with greater initiative and enterprise to move out of mainstream, large-scale organizations where the culture of accountable performance prevails. Teachers move out of public schools to private schools and charter schools. Engineers move out of large corporations to boutique firms. Enterprising government employees become consultants. There is a healthy element in this. But surely the large-scale organizations of our society are the poorer for driving out those most likely to innovate and initiate. The more that work becomes a matter of filling in the boxes by which performance is to be measured and rewarded, the more it will repel those who think outside the box.

Costs to prioductivity! Economists who specialize in measuring economic productivity report that in recent years the only increase in total factor productivity in the American economy has been in the information-technology-producing industries.[11] A question that ought to be asked is to what extent the culture of metrics-with its costs in employee time, morale, and initiative, and its promotion of short-termism-has itself contributed to economic stagnation?

16

WHEN AND HOW TO USE METRICS

A CHECKLIST

There is nothing intrinsically pernicious about counting and measuring human performance. We all tend to project broad-ranging conclusions based on our inevitably limited experience, and measured data can serve as a useful counterpoint to those subjective judgments. The sort of measurements with which this book is concerned are performance metrics that quantify human achievement and failure. There are legitimate metrics of performance in almost every organization.

In our case studies, we've seen many instances in which metrics has been useful and effective.

In policing, computerized statistics of the incidence of crimes (Compstat) were used to good purpose, to discover where problems were greatest and where police resources were best deployed. It ran into problems only when officials used the threat of demotion or lack of promotion against those lower in the hierarchy to try to bring down the reported crime rates.

In universities, faculty evaluations can be enhanced by numerical data about publications and teaching. The metrics go awry when they are used mechanically by those who are not in a position to evaluate the accuracy and significance of the data.

In primary and secondary education, standardized tests can be used to inform teachers of how much or how little their students are learning in particular subjects. Teachers can consult with their colleagues, and adjust their methods and curriculum as a result. Problems arise when the tests become the primary basis on which teachers and schools are rewarded or punished.

In medicine, Peter Pronovost's Keystone project demonstrates how effective diagnostic metrics can be in lowering the incidence of medical errors, when what is measured accords with the professional values of practitioners. The success of the Geisinger medical system illustrates the remarkable improvements made possible by computerized measurement when integrated into an institutional culture based on cooperation, where the setting of measurement criteria and the evaluation of performance are done by teams that include physicians as well as administrators. In both cases, metrics were used in ways that appealed to intrinsic motivation and to professionalism. But elsewhere in the medical system, as we've seen, the use of reward for measured performance sometimes proved fruitless or led to perverse outcomes.

Reflections on the best use of performance metrics by the U.S. Army in its counterinsurgency campaigns showed that while standardized metrics are often deceptive, metrics developed to fit the specific case, especially by practitioners with local experience, could be genuinely informative. The challenge in such cases is to abandon universal templates and discover what is worth counting, and what the numbers actually mean in their local context.

As we've seen time and again, measurement is not an alternative to judgment: measurement *demands* judgment: judgment about whether to measure, what to measure, how

to evaluate the significance of what's been measured, whether rewards and penalties will be attached to the results, and to whom to make the measurements available.

Should you find yourself in a position to set policy, here are the questions you should ask, and the factors you should keep in mind, in considering *whether* to use measured performance, and if so, *how* to use it. They constitute a checklist of successful performance measurement. Given what we've said about the hazards of metric fixation, consider at every point that the best use of metrics may be not to use it at all.

THE CHECKLIST

1. What *kind* of information are you thinking of measuring? The more the object to be measured resembles inanimate matter, the more likely it is to be measureable: that is why measurement is indispensable in the natural sciences and in engineering. When the objects to be measured are influenced by the process of measurement, measurement becomes less reliable. Measurement becomes much less reliable the more its objtKt is human aotivity, since the objects-people-ane self:.conscious, and are capable of reacting to the process of being measured. And if rewards and punishments are involved, they are more likely to react in a way that skews the measurement's validity. By contrast, the more they agree with the goals of those rewards, the more likely they are to react in a way that enhances the measurement's validity.
2. How *useful* is the information? Always begin by reminding yourself that the fact that some activity is measureable does not make it worth measuring, indeed, the ease of measuring may be inversely proportional to the significance of what is measured. To put it another way, ask your-

self, is what you are measuring a proxy for what you really want to know? If the information is not very useful or not a good proxy for what you're really aiming at, you're probably better off not measuring it.

3. How useful are *more* metrics? Remember that measured performance, when useful, is more effective in identifying outliers, especially poor performers or true misconduct. It is likely to be less useful in distinguishing between those in the middle or near the top of the ladder of performance. Plus, the more you measure, the greater the likelihood that the marginal costs of measuring will exceed the benefits. So, the fact that metrics is helpful doesn't mean that more metrics is more helpful.
4. What are the costs of *not* relying upon standardized measurement? Are there other sources of information about performance, based on the judgment and experience of clients, patients, or parents of students? In a school setting, for example, the degree to which parents request a particular teacher for their children is probably a useful indicator that the teacher is doing something right, whether or not the results show up on standardized tests. In the case of charities, it may be most useful to allow the beneficiaries to judge the results.
5. To what *purposes* will the measurement be put, or to put it another way, *to whom* will the information be made transparent? Here a key distinction is between data to be used for purposes of internal monitoring of performance by the practitioners themselves versus data to be used by external parties for reward and punishment. For example, is crime data being used to discover where the police ought to deploy more squad cars or to decide whether the precinct commander will get a promotion? Or is a surgical

team using data to discover which procedures have worked best or are administrators using that same data to decide whether the hospital will be financially rewarded or penalized for its scores? Measurement instruments, such as tests, are invaluable, but they are most useful in internal analysis by practitioners rather than for external evaluation by public audiences who may fail to understand their limits. Such measurement can be used to inform practitioners of their performance relative to their peers, offering recognition to those who have excelled and offering assistance to those who have fallen behind. To the extent that they are used to determine continuing employment and pay, they will be subject to gaming the statistics or to outright fraud.

Remember that, as we've seen, performance metrics that link reward and punishment may actually help reinforce intrinsic motivation when the goals to be rewarded accord with the professional goals of the practitioners.[1] If, on the other hand, the scheme of reward and punishment is meant to elicit behavior that the practitioners consider useless or harmful, the metrics are more likely to be manipulated in the many ways we've explored. And if the practitioners are too geared toward extrinsic reward, they may well react by focusing their activity on what is measured and rewarded, at the expense of other facets of their work that may be equally important. For all these reasons, "low stakes" metrics are often more effective than when the stakes are higher.

Recall that direct pay-for-performance works best to the degree that people are motivated by extrinsic reward rather than intrinsic motivation, that is, when they care about making more money rather than about the other potential benefits of their work, social and intellectual.

That may be because they are in a field, such as finance, in which people measure their own vocational success almost entirely in terms of the amount they earn. (As we've noted, that doesn't preclude them from using their earnings for a wide range of purposes, including selfless ones.) It is when the job offers few other attractions-when it is repetitious and leaves little room for the exercise of choice, for example replacing windshields or preparing hamburgers-that pay for measured performance is more likely to work.

6. What are the *costs* of acquiring the metrics? Information is never free, and often it is expensive in ways that rarely occur to those who demand more of it. Collecting data, processing it, analyzing it-all of these take time, and their expense is in the opportunity costs of the time put into them. To put it another way, every moment you or your colleagues or employees are devoting to the production of metrics is time not devoted to the activities being measured. If you're a data analyst, of course, producing metrics *is* your primary activity. For everyone else, it's a distraction. So, even if the performance measurements are worth having, their worth may be less than the costs of obtaining them. Remember, too, that those costs in human time and effort are themselves almost impossible to calculate-another reason to err on the side of caution.
7. Ask why the people at the top of the organization are demanding performance metrics. As we've noted, the demand for performance measures sometimes flows from the ignorance of executives about the institutions they've been hired to manage, and that ignorance is often a result of parachuting into an organization with which one has little experience. Since experience and local knowledge

matter, lean toward hiring from within. Even if there is someone smarter and more successful elsewhere, their lack of particular knowledge of your company, university, government agency, or other organization may not outweigh the benefits of hiring from within.

a. *How* and *by whom* are the measures of performance developed? Accountability *metrics* are less likely to be effective when they are imposed from above, using standardized formulas developed by those far from active engagement with the activity being measured. Measurements are more likely to be meaningful when they are developed from the bottom up, with input from teachers, nurses, and the cop on the beat. That means asking those with the tacit knowledge that comes from direct experience to provide suggestions about how to develop appropriate performance standards.[2] Try to involve a representative group of those who will have a stake in the outcomes.[3] In the best of cases, they should continue to be part of the process of evaluating the measured data.

Remember that a system of measured performance will work to the extent that the people being measured believe it is worth it. So far, in this chapter, we've taken the perspective of those in a position to decide whether and how to institute metrics. But what if you are not in such a position, if you're further down in the organizational hierarchy, where you are expected to execute metrics-a mid-level manager, say, or the head of an academic department? Then, you face a choice. If you believe in the goals for which the information is being collected, then your challenge is to provide accurate data in the most efficient way possible, one that demands the least time of you and those you manage. If, by contrast, you believe that the goals

are dubious and the process wasteful, you might try to convince your superiors of that (perhaps by giving them a copy of this book). If that fails, then your task is to provide data in a way that takes the least time, meets minimal standards of acceptability, and won't harm your unit.

If you're near the ***top*** of the organization, making decisions about metrics, reread the previous paragraph, keeping in mind the different ways in which those below you might react. Metrics wo.rks best when those measured buy into its purposes and validity.[4]

9. Remember thauu.en ***the best measures are. s1,bject to corruption or goal diuersion.*** lnsofar as individuals are agents out to maximize their own interests, there are inevitable drawbacks to all schemes of measured reward. If, as is currently still the case, doctors are remunerated based on the procedures they perform, that creates an incentive for them to perform too many procedures that have high costs but produce low benefits. But pay doctors based on the number of patients they see, and they have an incentive to see as many patients as possible, and to skimp on procedures that are time-consuming but potentially useful. Compensate them based on successful patient outcomes, and they are more likely to cream, avoiding the most problematic patients.[5]

That doesn't mean that performance measures should be abandoned just because they have some negative outcomes. Such metrics may still be worth using, despite their anticipatable problems: it's a matter of trade-offs. And that too is a matter of judgment.

10. R member that sometimes, recogni2Jing the limits 0£ the possiblejs the beginning of wisdom. Not all problems are soluble, and even fewer are soluble by metrics. It's not true

that everything can be improved by measurement, or that everything that can be measured can be improved. Nor is making a problem more transparent necessarily a step to its solution. Transparency may make a troubling situation more salient, without making it more soluble.

In the end, there is no silver bullet, no substitute for actually knowing one's subject and one's organization, which is partly a matter of experience and partly a matter of unquantifiable skill. Many matters of importance are too subject to judgment and interpretation to be solved by standardized metrics. Ultimately, the issue is not one of metrics-versus judgment, but metrics as informing judgment, which ►ncludes knowing how much weight to gi_ve to metrics,:recogni;z;ing their characteristic distoritions, and appreciating what can't be.measured. In recent decades, too many politicians, business leaders, poli$_{cy}$ makers, and academic officials have lost sight of that.



Board Ends Work Session

Dave VanLoo, Ph.D.

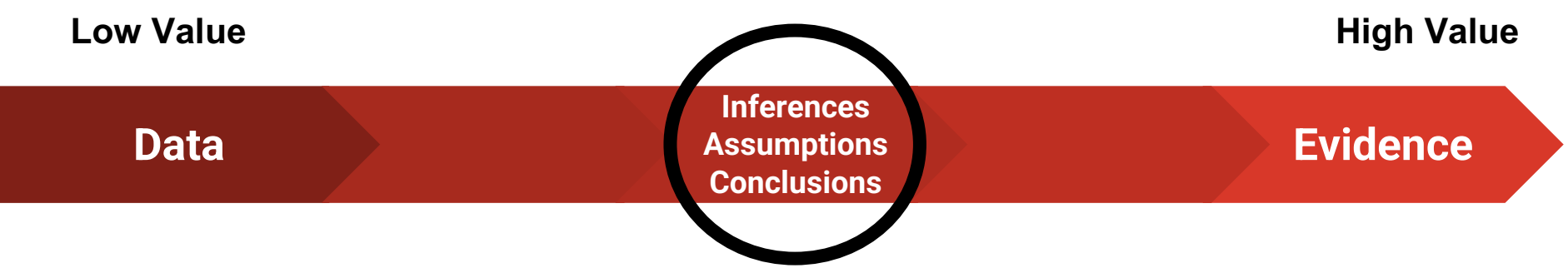
Est. 1883
BEND LAPINE
Schools
EDUCATING THRIVING STUDENTS

Low Value

High Value

Data are simply factual information such as numbers, percentages, and statistics.

Evidence is data that is relevant and furnishes proof that supports a conclusion.



The appropriateness of the inferences, assumptions, and conclusions people make based on the data is central to effectively using data and impacts the quality of the evidence.

Do the proposed metrics provide clear ***evidence of schools' impact*** in advancing the corresponding Board Ends?

Examples of inferences, assumptions, or conclusions people may draw based on the data used as evidence of Board Ends

- **Direction**: Should the data be increasing, decreasing, or unchanging?
- **Level**: What is the target goal and how close is the district to meeting it?
- **Reliability & Validity**: Are the data a high quality measure of the End?
- **Action / Impact**: Do the data point to specific actions that drive meaningful improvement?
- **Scope**: Does this apply to all staff, students, and families or only some?
- **Values**: What does our choice of data say about district values and beliefs?

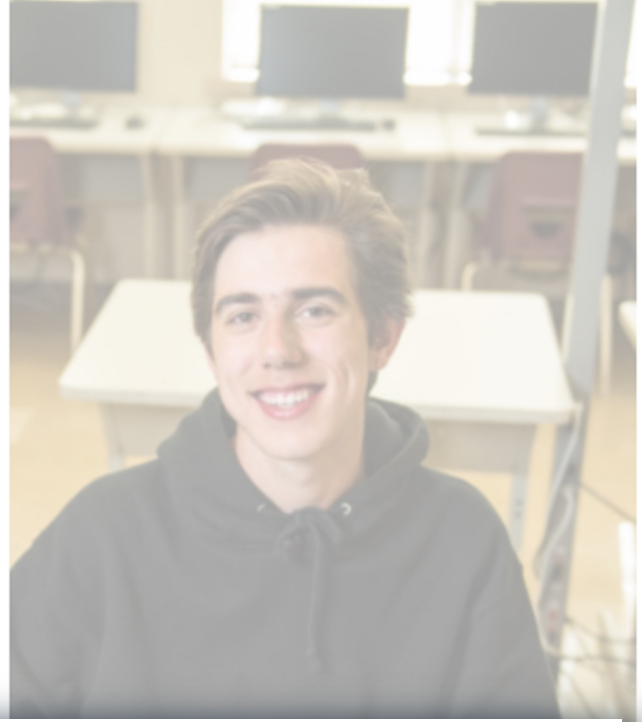
OSAS Test Scores

- **Direction**: Should OSAS scores be increasing, decreasing, or unchanging?
- **Level**: What percentage of students should be meeting standards on OSAS?
- **Reliability & Validity**: Are OSAS scores a high quality measure of the academic excellence?
- **Action / Impact**: Do OSAS scores help inform specific actions that drive meaningful improvement?
- **Scope**: Do we care about OSAS scores for all students?
- **Values**: What does our use of OSAS scores say about district values and beliefs?

Board Ends #’s 1-3

1. Students develop a **strong academic foundation**
    - Measures of academic success from kindergarten to twelfth grade.
2. Students have a **passion, purpose, and plan** for their future
    - Culminating measures of preparedness for success in life beyond K-12.
3. Students, families, and staff **experience wellness, inclusion, and belonging** in our schools
    - Measures of school climate (the way the school feels) and culture (the way we do things at this school-shared beliefs and values).

- Equity (focus on HU groups) and engagement are inherent throughout all three ends.
- **The evidence used to measure progress on the Ends should be things over which schools have substantial control!**


Sesión de Trabajo de las Metas de la Mesa Directiva

Dave VanLoo, Ph.D.

Est. 1883
BEND LAPINE Schools
Educating Thriving Students

Valor Bajo

Valor Alto

Datos

Inferencias
Suposiciones
Conclusiones

Evidencia

Datos son simplemente información fáctica, como números, porcentajes y estadísticas.

Evidencia son datos que son relevantes y proporcionan pruebas que respaldan una conclusión.

Valor Bajo

Valor Alto

Lo apropiado de las inferencias, suposiciones y conclusiones que las personas hacen con base en los datos es fundamental para usar los datos de manera efectiva e impacta la calidad de la evidencia.

¿Las métricas propuestas brindan ***evidencia clara del impacto de las escuelas*** en el avance de las metas correspondientes de la mesa directiva?

Ejemplos de inferencias, suposiciones o conclusiones que las personas pueden sacar en función de los datos utilizados como evidencia de las metas de la mesa directive

- **Dirección:** ¿Los datos deberían ser crecientes, decrecientes o invariables?
- **Nivel:** ¿Cuál es el objetivo a alcanzar y qué tan cerca está el distrito de alcanzarlo?
- **Confiabilidad y Validez:** ¿Son los datos una medida de alta calidad de la meta?
- **Acción/Impacto:** ¿Los datos apuntan a acciones específicas que impulsan una mejora significativa?
- **Alcance:** ¿Esto se aplica a todo el personal, estudiantes y familias o solo a algunos?
- **Valores:** ¿Qué dice nuestra selección de datos sobre los valores y creencias del distrito escolar?

OSAS Puntajes de las pruebas

- **Dirección:** ¿Deberían los puntajes de OSAS aumentar, disminuir o permanecer inalterables?
- **Nivel:** ¿Qué porcentaje de estudiantes debería cumplir con los estándares de OSAS?
- **Confiabilidad y Validez:** ¿Son los puntajes OSAS una medida de alta calidad de la excelencia académica?
- **Acción / Impacto:** ¿Los puntajes OSAS ayudan a informar acciones específicas que impulsan una mejora significativa?
- **Alcance:** ¿Nos preocupamos por los puntajes OSAS para todos los estudiantes?
- **Valores:** ¿Qué dice nuestro uso de las puntuaciones OSAS sobre los valores y creencias del distrito escolar?

Metas de la Mesa Directiva # 1-3

1. Los estudiantes desarrollan una **base académica sólida**
    - Medidas de éxito académico desde el jardín de niños hasta el duodécimo grado escolar.
2. Los estudiantes tienen una pasión, un propósito y un plan para su futuro
    - Medidas culminantes de preparación para el éxito en la vida más allá de los grados K-12.
3. Los estudiantes, las familias y el personal escolar **experimentan bienestar, inclusión y pertenencia** en nuestras escuelas
    - Medidas del clima escolar (la forma en que se siente el estar en la escuela) y la cultura (la forma en que hacemos las cosas en esta escuela: creencias y valores compartidos).

- La equidad (enfocarse en los grupos históricamente menos atendidos) y el compromiso son inherentes a las tres metas.
- **¡La evidencia utilizada para medir el progreso en las metas deben ser cosas sobre las cuales las escuelas tengan un control sustancial!**

BEND-LA PINE SCHOOLS
Administrative School District No. 1
Deschutes County, Oregon
GOVERNANCE POLICY

Name: Board Ends
Section: Governance Policy
Code: Ends/Results

PROMISE

Every student in Bend-La Pine Schools is known by name, strengths, and needs, and graduates prepared for college, career, community engagement, and life.

GOALS

Outcomes and Experiences

1. Students are engaged and develop a **strong academic foundation** as measured by the following, overall and for historically underserved subgroups:
    - (1a) Mastery of ELA & Math foundational knowledge and skills by the end of 1st Grade, as measured by standardized assessments
    - (1b) ELA, Math, & Science proficiency rates in 3rd-8th, as measured by the Oregon Statewide Assessment System (OSAS)
    - (1c) ELA and Math growth rates in 4th-8th grades as measured by the Oregon Statewide Assessment System (OSAS)
    - (1d) The percent of 9th graders on track for graduation, as measured by credit attainment
    - (1e) The percent of students designated as English learners that are on track to acquire English proficiency, as measured by Oregon's English Language Proficiency Assessment (ELPA)
    - (1f) Efficacy of academic program, as measured by surveys and/or focus groups
2. Students have a **passion, purpose, and plan** for their future as measured by the following, overall and for historically underserved subgroups:
    - (2a) 4-year and 5-year graduation rates, 5-year completer rate
    - (2b) The percent of graduates who earn a diploma plus complete at least one of the following career and life indicators:
        i. 2+ credits of Advanced Placement (AP) or International Baccalaureate (IB)
        ii. 2 semester or 3 quarter hours of college credit eligible coursework
        iii. State Seal of Biliteracy
        iv. 2 years of Junior ROTC
        v. CTE Concentrator (2+ credits in a program)
        vi. Meets college readiness benchmarks for language arts and mathematics on high school OSAS, ACT, or SAT
        vii. Meets full admission requirements for all Oregon Public Universities.
    - (2c) Students' preparedness for their future, as measured by surveys and/or focus groups
    - (2d) ~~The percent of students who enroll in a two- or four-year college or university, as measured by National Student Clearinghouse data~~

3. Students, families, and staff **experience wellness, inclusion, and belonging** in our schools as measured by the following, overall and for historically underserved subgroups:
    - (3a) Bias incident data and trends
    - (3b) Student, family, and staff experiences of key elements of school culture (voice, belonging, and emotional/psychological wellness), as measured by survey and/or focus groups
    - ~~(3c) Chronic absenteeism rate~~
    - ~~(3d) Suspension rate~~

Inferences, Assumptions, and Conclusions in our Data

Directions

Refer to the document containing a draft of Board Ends #'s 1-3 and associated data metrics that are intended to provide evidence of schools' impact. The objective of this activity is to assess how well the proposed metrics measure schools efficacy in advancing each end.

The objective is not to assess whether the proposed metrics are something the district *should* monitor. All metrics listed contain important information that the district *should* monitor. The key question is whether each metric is a *high-quality, systems-level measure of the actions school and district staff take to support progress* in that specific Board End.

In the following pages, for each Board End, address the three key questions. For Ends #2 and #3, also consider the inferences and evidence provided for the metrics of college-going rates, student attendance, and student suspensions.

End #1 Pairs
Melissa & Marcus, Carrie & Amy, Shirley & Shimiko

End #2 Pairs
Marcus & Carrie, Amy & Shirley, Shimiko & Melissa

End #3 Pairs
Amy & Melissa, Shimiko & Carrie, Marcus & Shirley

Inferences, Assumptions, and Conclusions in our Data

End #1: Possible Inferences, Assumptions, and Conclusions - Schools Build a Strong Academic Foundation

- Available standardized assessments designed for systems accountability are reasonably good indicators of academics through middle school. These data help districts identify both strengths and areas where improvement is needed. Data also provides families with information on their own child's academic achievement in comparison to students at their school and the district at large.
- Equitable academic excellence is central to the mission of public education.
    - "If students attend schools that do not foster in them excellence in reading, writing, science, and math, and therefore leave them unprepared to achieve excellence and leadership in their chosen field, we have not created a more socially just world, no matter how committed to action we may be. Equity starts with achievement." (Doug Lemov, Teach Like a Champion 3.0)
- A desired outcome is continuous improvement and increasing levels of learning for all students.
- The academic preparation necessary to thrive in postsecondary education, careers, community engagement, and life begins early and requires years to accomplish.
- Stakeholder perception data of the district's academic programs is essential to measuring this End.

Key Questions

1. Are these inferences, assumptions, and conclusions we hope stakeholders will draw? What other inferences, assumptions, and conclusions can people draw from these data, and is that desirable?
2. Do the proposed metrics provide clear, high-quality evidence for *schools' efficacy* in supporting this Board End? If yes, how? If not, why not?
3. Should any metrics be added or removed? Proposals for change should answer two questions:
    - "What will be better if the changes are made?"
    - "What will be worse if the changes are made?"

Use the space below for writing key notes, thoughts, and ideas

Inferences, Assumptions, and Conclusions in our Data

End #2: Possible Inferences, Assumptions, and Conclusions - Schools Develop Students with a Passion, Purpose, and Plan their Future

- A high school diploma is a key K-12 outcome and is a minimum standard most students need for future success.
- For many students, future readiness requires preparation and experiences beyond the minimum standards of a high school diploma.
- There are many ways a diverse population of students can demonstrate a high degree of preparedness for life beyond high school.
- Stakeholder perception data are essential to measuring this Board End.

Key Questions

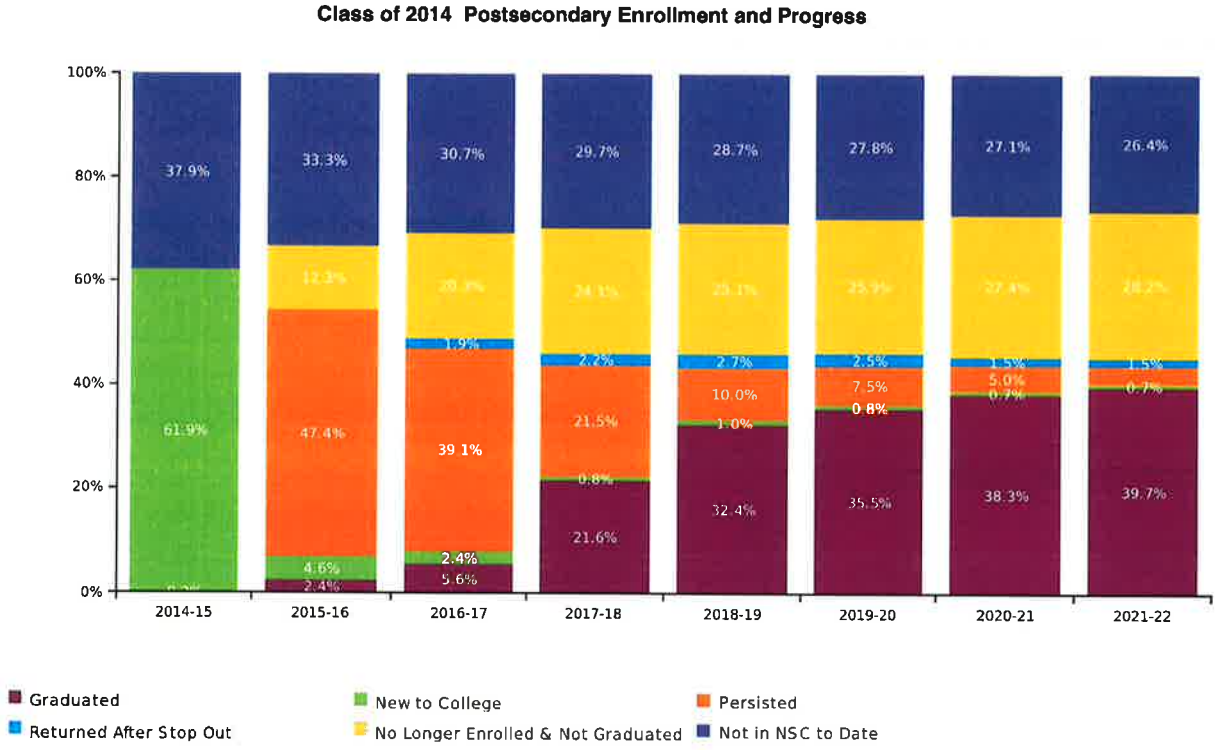
1. Are these inferences, assumptions, and conclusions we hope stakeholders will draw? What other inferences, assumptions, and conclusions can people draw from these data, and is that desirable?
2. Do the proposed metrics provide clear, high-quality evidence for *schools' efficacy* in supporting this Board End? If yes, how? If not, why not?
3. Should any metrics be added or removed? Proposals for change should answer two questions:
    - "What will be better if the changes are made?"
    - "What will be worse if the changes are made?"

Use the space below for writing key notes, thoughts, and ideas

Additional Notes on 2- and 4-Year College Going Rates Metric

What are some potentially problematic inferences, assumptions, and conclusions about this metric?

- Whether students enroll in college after graduation is mostly a function of what schools do.
- If a student enrolls in college after graduation, that represents clear success for both the district and the student.
- College is the preferred outcome for most students after high school.
- The percentage of students attending college should be increasing.



Inferences, Assumptions, and Conclusions in our Data

End #3: Possible Inferences, Assumptions, and Conclusions - Schools Create Safe, Welcoming, and Inclusive Environments for Students, Families, and Staff

- Effective, equitable schools actively work to monitor and reduce incidents that undermine a safe, welcoming, and inclusive climate for all stakeholders, and take proactive actions to create such a climate.
- Stakeholder perception data are essential to measuring this Board End. Students, families and staff all are key stakeholders.
- Thriving staff are essential to student success.

Key Questions

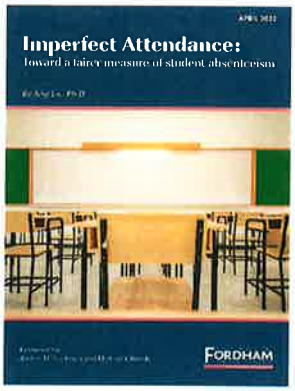
1. Are these inferences, assumptions, and conclusions we hope stakeholders will draw? What other inferences, assumptions, and conclusions can people draw from these data, and is that desirable?
2. Do the proposed metrics provide clear, high-quality evidence for *schools' efficacy* in supporting this Board End? If yes, how? If not, why not?
3. Should any metrics be added or removed? Proposals for change should answer two questions:
    - "What will be better if the changes are made?"
    - "What will be worse if the changes are made?"

Use the space below for writing key notes, thoughts, and ideas

Additional Notes on Chronic Absenteeism / Student Attendance Metric

What are some potentially problematic inferences, assumptions, and conclusions about this metric?

- Schools with lower attendance rates are less effective at building culture and climate that supports wellbeing, inclusion, and belonging
- Student attendance measures a similar construct from kindergarten through 12th grade.
- Student attendance measures a similar construct across different demographic groups.
- Schools have a great deal of control over student attendance.
- Student attendance rates, as currently reported, are a reliable and valid indicator of school climate, culture, and quality.



https://fordhaminstitute.org/national/research/imperfect-attendance-toward-fairer-measure-student-absenteeism (April 2022)

A focus during the post-Covid education recovery phase, then, should be making sure students return to school. Yet our systems for measuring their attendance—and holding schools accountable for getting kids back into classrooms—are woefully inadequate.

First, most jurisdictions rely exclusively on raw attendance rates and/or chronic-absenteeism rates, both of which are highly correlated with student demographics and other factors that schools generally cannot control.[5] Nevertheless, such metrics are ubiquitous in state accountability systems, with at least thirty states and the District of Columbia having adopted student absenteeism, chronic absenteeism, or variants thereof as a "measure of school quality" under the Every Student Succeeds Act.

Second, many states and districts do a poor job of measuring attendance because they can't (or choose not to) differentiate between full-day absences and partial-day ones (as in, when students show up for some classes but not for others). Prior research shows that partial-day absenteeism in secondary school is rampant, mostly unexcused, and explains more missed classes than full-day absenteeism.[6] Part-day absences increase with each middle school grade and then rise dramatically at the transition to high school.

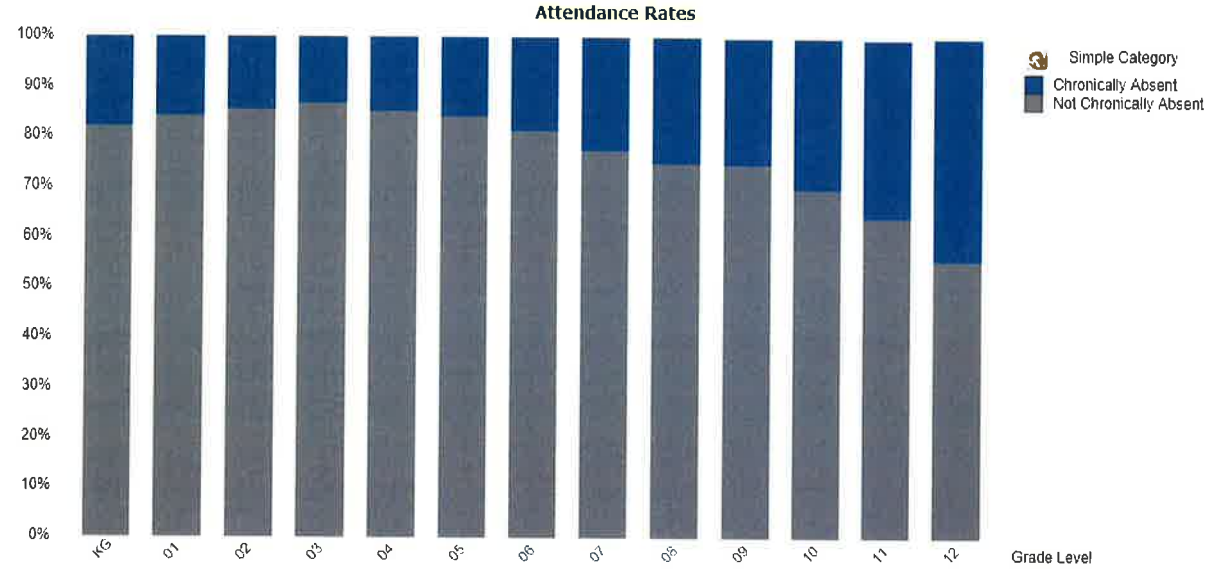
In short, the most widely adopted "fifth indicator" under ESSA has been framed in ways that are hopelessly broad and unfair. But why? After all, collecting detailed attendance data ought to be

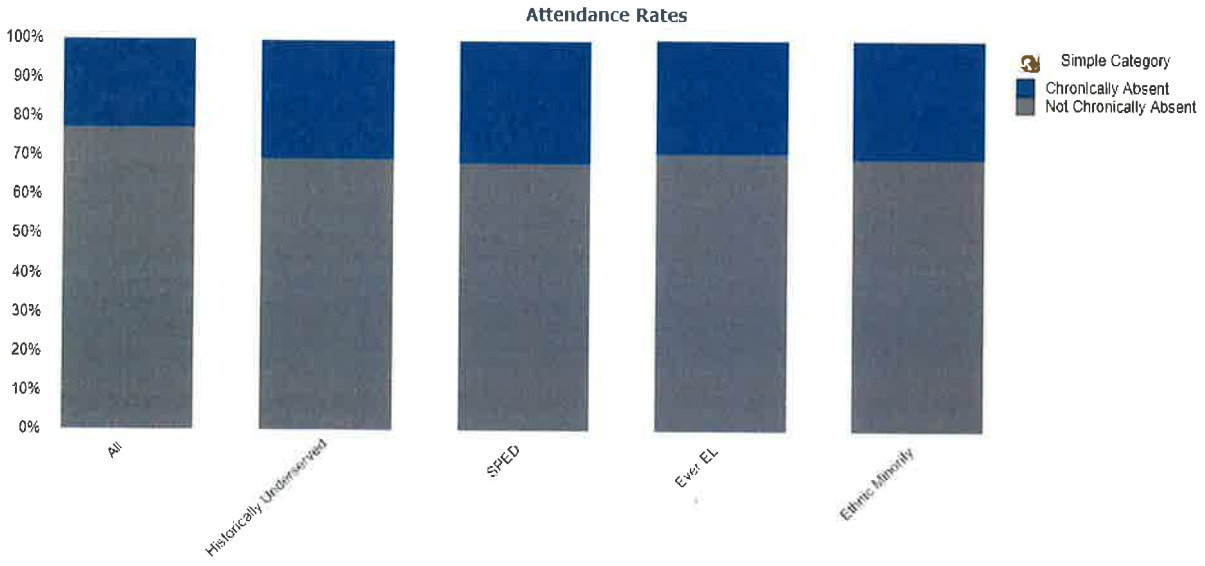
Implications

1. **Using raw student-absenteeism measures in accountability systems likely imparts credit or penalty to schools that don't deserve it.**

As the findings make clear, some schools do a better job of getting students to come to class regularly. Yet many of these schools are overlooked by states' existing accountability systems—even as others are rewarded for habits and circumstances that students developed before their arrival. To be clear, there are good reasons to report schools' raw attendance and chronic-absenteeism rates, which are transparent, parent-friendly, and crucial to understanding the depth of the problem. But if the goal is to hold schools or educators accountable for things they can control or help parents understand if their child's attendance is likely to improve, then raw attendance rates are unfair and uninformative (and the same is also true of *any* non-test-based indicator that fails to account for the things *students* bring to school).

Bend-La Pine Regular Attender / Chronic Absenteeism Data Since 2010-11





Additional Notes on Student Suspension Metric

What are some potentially problematic inferences, assumptions, and conclusions about this metric?

- Suspension rates, which typically are based on a small number of students, are always a valid and reliable measure of a school's culture and climate.
- Schools with better climates and cultures will have lower suspension rates. Schools with worse cultures and climates will have higher suspension rates.
- District-level suspension rate data provide an accurate view of how suspensions are used at the school level.
- Differential suspension rates are a problem in Bend-La Pine.
- Suspension typically is an inappropriate response to student misbehavior.

Raw Counts of Bend-La Pine Secondary Student Suspensions Since 2013-14