DISABLING NUMBERS:

ON THE SECRET CHARM OF NUMBERESE AND WHY IT SHOULD BE RESISTED

Anna Sfard

Institute of Education, University of London, UK & Michigan State University, US

Testing and measurement have always been inextricable parts of the processes of teaching and learning, but never did they occupy as central a place in educational processes as they do today. This unprecedented assessment frenzy is just a particular manifestation of a more general phenomenon: of our present tendency for speaking in numbers about absolutely anything, whatever the nature of the things that are talked about. Numerical discourses – the *numberese*, as they are collectively called on these pages – and their educational uses are the focal theme of this chapter. After illustrating the claim about the present prominence of numberese, I take a critical look at the reasons and effects of its popularity: I locate those properties of numerical discourses that appeal to educational decision makers, analyze underlying discursive mechanisms and argue that some of those features of numberese that make it decision-makers' favorite are a mere byproduct of certain discursive devices, and are thus a matter of appearances rather than of any genuine asset. In some extreme cases, such delusions may cause tangible damage. In conclusion I posit that immoderate, uncontrolled use of "numberese" with reference to people and their actions is detrimental rather than helpful to the educational enterprise.

1. The Hegemony of Numberese

The point of departure for the claims I wish to make in this chapter is the general assumption that one cannot escape if one takes seriously Vygotsky's insights on vo-devopment of language and thinking (Vygotsky, 19871987) and Wittgenstein's criticism of thinking-language dichotomy (Wittgenstein, 1953): Human thinking is the individualized form of interpersonal communication and the way we communicate with others and with ourselves has therefore a major impact on how we act and how we perceive the world (Sfard, 1995, 2005, 2008; cf. Edwards & Potter, 1992; Harré & Gillett,1995). Various types of communication, or simply discourses, differ from one another in their vocabularies and word use, in their mediators and routines, and in the narratives they produce. In particular, participants of each discourse have their own ways of constructing, defending, and eventually endorsing stories about the world. Among the most prominent and influential of these narratives are those we call *identities* – stories about who we are, with whom we belong, and what position we occupy among those who constitute our human environment. When discourses change, the whole world changes and our identities change with it. In particular, the human world described in terms of numbers is definitely not the same as the world described in number-free language.

Although numerical talk has always been around, its current ubiquity is without precedent. Our discourses, whether spoken or written, are saturated with

quantitative expressions. Numerical symbols are an inextricable element of any artifact and of any public space, and numerical talk can be heard whenever people open their mouth. Always part and parcel of discourses of natural sciences and economy – these discourses would simply not exist without the language of quantities – the numberese may be not an immediately obvious choice when it comes to humans. Still, we do employ quantitative terms also while talking about people, and we do so whether we describe individuals or collectives, and whether we refer to persons' physical properties or to the qualities of their actions. At a closer look, numbers are the principal ingredient of our identities: In addition to our bodily dimensions (which attract much more attention these days than they did in the past), we identify ourselves by test scores, exam results, final grades, intelligence quotients, ranks and levels, income, socio-economic indexes, states of possession, risks of genetic maladies, sports records, energy levels, placements on waiting lists, popularity ratings, and so on, and so forth. Since our capacity for measuring and our quantitative imagination are unbounded, there is no reason why this litany should ever end.

Add to this the omnipresent measuring tools that became available with the advent of information technology, and everything about us becomes quantifiable. Just to make my point, let me pause for a moment and take a quick look around… I enter the Internet and, at random, I find the following piece of news:

> [An economist], with her students as research assistants, staked out eight coffee shops in the Boston area and watched how long it took men and women to be served. Her conclusion: Men get their coffee 20 seconds earlier than do women. (There is also evidence that blacks wait longer than whites, the young wait longer than the old, and the ugly wait longer than the beautiful…..)[1]

In my email inbox I then spot a questionnaire with the help of which a group of researchers whose names sound vaguely familiar is trying to find out what should be considered as necessary ingredients of mathematics teacher preparation. 'We request no more than fifteen minutes of your time', they say in the letter. And indeed, they are using a Likert scale, so all I have to do is to tick numbers in boxes (and nobody seems bothered about the time I may need to think about my choices!). When I shift my eyes to the desk, I see a pile of student papers waiting for grading. I then notice a request for recommendation that arrived a few days earlier. A colleague is up for promotion and I am asked to assess his "productivity level," the "impact factor" of the journals in which he publishes his work, and his "professional standing in comparison to scholars at a similar stage in their careers." Eager to do my work as a referee on time, I am tempted to look at his *GoogleScholar* scores. Another tool to be found on Internet, called *Publish or Perish*, allows me, in a matter of seconds, to extract several additional measures: cites/year, cites/paper, cites/author, papers/author, etc. Above all, I can now help myself with new indices – h-index, g-index, hc-index, hI-index, and several others,

[1] http://www.slate.com/id/2177697/nav/tap3, retrieved on 15 Nov 07

all of them touted as being directly indicative of the quality of scholars' work. And although I am not sure of how the new indices are calculated and why they are supposed to reflect what people have in mind when speaking about quality, I now have an instrument with which to compare the candidate to everybody else in academia. And there are many other measures I can use: this person's grants, his scores in student surveys, the number of invited talks and keynotes he gave at different conferences – and the list is still long. Even if my instinct is to shy away from all these numbers, I do keep in mind that whatever I am going to say stands a better chance to be found truly convincing if I express it in numberese. After all, whoever knows a thing or two about the politics of discourses must realize that numerically-grounded statements are not anything one can easily argue with. They have the power to overwrite any other.

The steep rise in the popularity and dominance of numerical talk is explicable in view of the current proliferation of number-producing devices. The Internet alone is an inexhaustible supplier of brand-new numerical measures, only some of which constitute an answer to a pre-existing need for information, most of which are a mere byproduct of the very possibility of measurement. Many of the things that are now accorded with numerical labels were never before considered as in any way number-related. As will be instantiated below, some of the objects implied in the measures did not even exist before the numerical epithets brought them into being. But even if stronger than ever, our tendency for numerical labelling is not new. I am now going to argue that the attractiveness of the

numberese in any context, but in educational context in particular, stems, above all, from its four ostensible strengths: its informativeness, generality, rigor and objectivity. These are the properties that give it an appearance of the best way of talking for those interested in quick and effective decision making.

2. Educational Uses of Numberese: Why, How, and to What Effect

The excerpts in Box 1 are typical snippets of educational numberese, that is, of texts made possible by the (discursive) activity of quantitative educational assessment. I will use these excerpts to instantiate the four properties of this latter discourse and to substantiate my skepticism about its helpfulness. You are advised to examine the examples carefully before continuing reading.

Box 1: Examples of text produced as a result of quantitative educational assessment.

A. From a written testimony of University of Haifa undergraduate, recalling her experience as mathematics pupil in middle school[2]:

After studying with a friend for a whole week, hours after hours, I got the "amazing" grade: 18 [out of 100]. That very day I told my teacher that this was the last time he had seen me in his classroom. But the teacher told me that if I only tried, I'd perhaps be able, at the end of the year, to get 100, and certainly not less than 80. I finished my mathematics studies at level 3 [the lowest], but with the grade 96.

[2] Collected in 1999 and translated from Hebrew by the author.

B. From a job advertisement in *Work* supplement to the British daily *Guardian*[3]

To qualify for the scheme, you'll need a minimum 2.2 degree in any discipline.

C. From *IEA's TIMSS 2003 International Report on Achievement in the Mathematics Cognitive Domains*[4]

At the eighth grade, led by Singapore, 24 countries and the four benchmarking participants had achievement in the applying domain significantly higher than the international average. Romania, Bulgaria, Norway, and Serbia performed no differently than the international average and 18 countries performed significantly below this average. At the fourth grade, also led by Singapore, 14 countries and the US state of Indiana had achievement significantly higher than the international average, two countries (Italy and Australia) and the two Canadian provinces had achievement similar to the international average, and 9 countries had achievement below it.

D. From *Review*, the supplement to the British daily *Guardian*[5]

Mary Beard vividly remembers a day in her first year at Newnham College, Cambridge, when one of her friends saw a marked essay lying on her desk. He picked it up and read the tutor's comment: 'This is very good; I think it

[3] 10 Nov 07, p. 24

[4] In the section titled "Applying Knowledge and Conceptual Understanding," p. 23 http://books.google.com/books?id=0oaXBqUgLyQC&pg=PA159&lpg=PA159&dq=svein+lie+timss&source=web&ots=aeZRVIlhGK&sig=gtpRwXPV09LZW1HC_hHdQZFLP-A#PPA138,M1 retrieved on 15 Nov 07 on

[5] 5 Nov 07, p. 11

would get the first.' 'You,' he spluttered, 'get a first?' 'Even in the mid-70s,' Beard recalls, there were 'lots of men who thought that women were destined only to get 2.1s.… 'From that moment,' she laughs, 'I was bloody determined to show them.' And she has shown them: Beard is now a professor in Cambridge and the best known classicist in Britain.

2.1 Informativeness

We may be all too familiar with the kind of texts presented in the box to notice how tightly packed with information they are. Just think about the complexity hiding beneath the simple idea of grade. Take, for instance, the 18 and 96 in excerpt A, the 2.2 in excerpt B, or the 'first' and 2.1 in excerpt D. All these numbers are but tips of icebergs; they are all end products of intricate discursive processes of differing length and depth: 18 and the 'first' have been produced through the evaluation of a particular piece of student work, 96 summarizes several basic evaluations of this earlier type, the 2.2 in the job advertisement refers to the final assessment supposedly combining all the grades accorded to the student through numerous years of study. Be they as stuffed with a hidden discourse as they are, individual grades are nowhere near in their tacit complexity to the *international average* in excerpt C. This time, hundreds of thousands, if not millions, of individual grades for individual test items had to be successively produced before they could be combined into test grades; the test grades were then brought together to yield national scores and averages; to arrive at the international

average, this last set of numerical data had to be subjected to yet another intricate, multistep discursive procedure.

The single number in excerpt C that epitomizes uncountable discursive actions of millions of students, evaluators, and data analysts, is a product of repetitive reifying – of replacing lengthy texts (e.g. student's written answers to test questions) with numbers, applying new discursive processes to these numbers (as in the process of calculating the total grade for the test), replacing the new processes with new numbers, and so forth. Halliday (1987) calls the texts resulting from such recurrent reifications *synoptic*. In such texts, 'the world is a world of things, rather than of happenings; of product, rather than of process; of being rather than becoming' (pp. 146-47).

The mechanism of reifying is invaluable in scientific discourse, where it is used to describe the functioning of agentless objects, but it may be less than helpful when applied to people and their actions. Here, the price of thus earned communicational thriftiness and intensity may be too high to afford. First, the repetitive reifications lower the resolution of the resulting picture, often to such degree that it becomes impossible to translate the assessment into a truly useful practical advice. Indeed, a single grade ties together thousands upon thousands of individual activities. In the numerical images of reality, people are only as different as their numerical labels allow them to be. The users of the numerical information are thus habitually overlooking the potentially significant differences

hiding beneath uniform numerical surfaces. Second, in successive reifications the discursive history of the numbers is irreversibly lost and unpacking the grades into what has actually been done by individual students and teachers in specific situations becomes virtually impossible. In view of this, trying to interpret the numerical scores is a hazardous enterprise. How to account for Singapore's consistently high scores? Are they a matter of teaching methods, genetic predispositions of the students, or some cultural idiosyncrasies, which make the success irreproducible in other cultures? (c.f Wang & Lin, 2005; Sfard & Prusak, 2005). Besides, did this success involve competences that we really wish our students to have? After all, the related processes of using mathematics in solving problems and of responding to math exam questions, although deceptively similar, are in fact two different types of discursive activity, oriented toward different goals, stimulated by different prompts and constituted by differently made choices (Lave 1988).

2.2 Generality

Another apparent reason for the attractiveness of numerical assessments is their wide applicability. Once an action or a series of actions is evaluated and accorded with a grade, we tend to see this numerical label as indicative of a pretty general property of the actor. Think, for example, about the requirement of “a minimum 2.2 degree in *any discipline*” in the job advertisement in Box 1 (excerpt B, emphasis added). Such requirement is only justified in the mouth of a person who considers the grade as independent of specificities of the learning process, and who

believes that a satisfactory score in one subject is a reliable predictor of satisfactory performance in any other type of activity. To put it differently, the advertiser treats the grade as a general property of a person rather than of the way this person has performed some specific tasks.

Although this kind of assumption seems empirically testable, we rarely feel any need for justification while letting it guide our decisions. Once again, certain discursive mechanisms may be responsible, at least in part, for this state of affairs. Consider the following sentence from a recent issue of Time Magazine:

'The average American child watches four hours of TV a day'[6]

At first sight, nothing seems strange here. On closer look, however, one realizes that the adjective *average*, which should be used as a descriptor *of a certain number* related to children's activities (one should rather say 'American children watch TV, in average, for four hours a day'), shifted to the noun *child* itself, thus creating a new entity, "an average child." The relocation of the term *average* implies the assumption that doing things in an "average way" comes in clusters: you watch TV the number of hours equal to the national average of watching hours – and you do all other things in an "average" way. Whereas not directly related to educational assessment, this example provides a perfect instantiation of one of the many linguistic transformations that underlie our belief in the generality

[6] TIME Magazine, 15 Oct 2007, section *Numbers*, p. 24

of messages carried by numerical labels and our resulting tendency for overgeneralizing these messages.

Once translated from a property of an action into a property of an actor, numerical evaluations are likely to cause temporal overgeneralization as well. Numbers accorded to a person for a particular performance stay the same as the performance goes on, and they are thus only likely to act as self-fulfilling prophecies. Grades and the resulting epithets such as *gifted* or *learning disabled*, although constructed on the basis of one's former actions, are usually read as statements about the subject's future. In result, low scores for past activities start perpetuating failure, whereas high grades are only too likely to beget further series of high grades. To see how this works, it is enough to look again at excerpt A in Box 1. The bad grade, 18 out of 100, would have acted as a trigger for a downfall, if not for a prompt preventive action on the part of her teacher (note, however, that whereas this action dissuaded the student from giving up entirely, it did not prevent her from aiming low!) Not all the students, however, are blessed with this kind of support. Only too often do I see in my studies how one or two local failures lead to a sudden avalanche and an irreversible damage to the student's further mathematical learning (see e.g. Ben-Yehuda et al., 2005). Let me stress that in the present context, the term *failure* does not refer just to one's inability to cope with certain types of problems, but also to the formal conferral of labels that automatically position the person at the very lowest end of the spectrum, in the category of "those who cannot." Note that this comparative aspect, made possible

by the use of numbers, played also a decisive role in Mary Beard's case (excerpt 4). Indeed, had the tutor satisfied himself with a qualitative, non-comparative appraisal, one that would have not implied her being *above* many of her classmates, most of the boys included, the effect would probably be not as dramatic as it was.

To sum up, numerical labels tend to stretch beyond the sets of things for which they were originally designed, thus dangerously extending claims about samenesses, equivalences and permanence. Generally believed to provide the best tools we have for describing reality, educational numberese is, in fact, the most underrated instrument for shaping the world around us. As long as we remain unaware of this latter fact, we mold reality with numbers unwittingly, and not necessarily the way we would like this reality to be.

2.3 Rigor

Numerical labels divide the world into clear-cut, non-overlapping categories and the resulting picture seems neat, precise, and free from uncertainty. In the discourse of numerical assessment there is little room for hedges and qualifiers and for modality other than full certainty: Either one has a certain grade or one hasn't; being in the "upper 10% of the class" or having a 2.2 degree is a yes-or-no affair. Once a grade was accorded, the owner of the grade – be it an individual student, as in excerpts A or D, or a country, as in C – has been positioned in an

unequivocal manner among all the other grade owners.[7] The neatness of numbers and their relations is bequeathed to the putative entities represented by the numbers. Thus, for example, once we start measuring "mathematical competency," we are also likely to declare that the student who scored 70 on the mathematics test is "less competent" or even "less able" than the one who scored 85. We rarely ask ourselves what mathematical competency *is* apart from the numbers with which it is measured.

And yet, when we do think about the nature of measured entities – learning, competence, understanding, intelligence, and so forth, we realize that the reality is much messier than its numerical image. Dazzled by the algorithmic rigor of numerical calculations we tend to forget that according grades to students' actions is a matter of imperfect, non-algorithmic human judgment. And indeed, the clean numerical picture of students' learning may be but a cover for messiness. Only too often, the discourse of quantitative assessment forces unruly, ill-defined segments of reality into the straightjacket of crystal-clear numerical categories the way Cinderella's stepsister's pushed their big feet into the glass shoe: by cutting off whatever does not fit into the slot. Moreover, there is little consistency in how we perform this cutting from one situation to another. This is certainly not the way to attain rigorous understanding of what is going on when students learn, nor is it a

[7] The language of positioning, by the way, is exactly the same whatever the nature of the graded entities: the expression "x performs below average" is equally useful when x refers to an individual and when it denotes a whole nation (see excerpt C again.) The preservation of the language in passing from individual entities to encapsulated aggregates of such entities is what helps in camouflaging the quickly growing complexity.

good way to collect reliable insights on which to base improvements of our own teaching practices.

2.4 Objectivity

All that has been said so far does not yet explain why we fall so easily a prey to the temptations of quantitative assessment. True, such features as informativeness, generality and rigorousness are all highly attractive, but what is the use of a discourse which only *pretends* to have all these advantages? The most obvious explanation for our weakness for numberese is that we are mostly unaware of the delusion. We take numbers bona fide, grateful for all they seem to be offering.

This confidence in numbers stems from yet another property of numerical statements: their appearance as mind-independent, world-imposed truths. Just look at the examples in Box 1: all of them speak about numerical assessments, but none of them mentions the assessor. The numbers and their interpretations are quoted as if they had no author. For instance, in excerpt C one reads that 14 countries have “performed below [international] average”, rather than being told that "assessors accorded numbers lower than international average to 14 countries." The matter-of-fact tone of the direct statement about "below average performance" implies that international average and the national grades are “objective” - they come from the world itself. As such, it is not up to humans to question the numbers or to try to change them. Similarly, in excerpt 4, people are

said to be “destined” to the grades they get, and even if stated with tongue in cheek, this assertion would not have resonated with Guardian readers if it didn’t build on some ‘genuine’ common belief in the direct relation between human beings and numbers.

Obliteration of the assessor is a highly consequential move. It results in what Bakhtin (1986) called monological discourse, a discourse whose narratives appear to be told in a single non-human voice – “‘the voice of the life itself,’ ‘the voice of nature,’ ‘the voice of people,’ ‘the voice of God,’ and so forth.” (p. 163). In monologically told stories no space is left for human agency – monological storytellers do not seem to consider the possibility that different people might have produced different numerical measures for the same phenomena. Unmediated by human beings, the relation between numbers and the things they measure becomes a part of the external reality. In short, it is the discursively engineered appearance of mind-independence and objectivity that fills us with confidence in numbers, blinds us to their pitfalls, and turns numberese into a powerful discourse, likely to overwrite any other.

3. Lesson: Beware of Educational Uses of Numberese

As argued in this chapter, all the ostensible advantages of quantitative assessments - their ostensible informativeness, generality, rigor and objectivity are simply byproducts of certain discursive techniques. Through these techniques, diversity is successfully disguised as uniformity and mere speculations acquire an

appearance of unassailable truths. When combined together, these two delusions become infallible sources of unhelpful diagnoses and potentially disastrous educational decisions. Ironically, therefore, the favorite tool of those who purport to be helping students may, in fact, be one of the greatest obstacles to students' learning.

Once we become aware of this, we may also realize that we can, and probably should, argue with numbers and seek alternative ways of picturing reality. This, however, may be easier said than done. In principle, educational assessment does not have to be numerical. Take, for example, assessment for learning (see e.g. Black et al., 2002; Cooper, 2006). If the aim is to give the student feedback about her progress, one obvious option is to analyze samples of her work, pointing to its specific strengths and weaknesses. In addition to its being free from the said weaknesses of numerical assessment, this type of reaction is probably much more effective: it provides the learner with specific, tailor-made information about what needs to be improved and how. However, non-numerical talk does not have the one special property of quantitative feedback that endears quantitative assessment to politicians and educational decision-makers: it does not support easy comparisons and does not order all the learners in a single line, where everybody can be juxtaposed to everybody else. Without such linear ordering, there is no divide between *normal* and *abnormal* or *healthy* and *pathological*, and one cannot speak in terms of *red lines*, *average*s, *tops* and *bottoms*. But all these latter expressions are exactly what decisions makers need to implement perhaps

the most important of their tasks: to make selections and orchestrate exclusions (unfortunately, our society is not yet organized well enough to afford being left without any of these.) Thus one cannot expect them to give up their favorite tool without struggle.

Still, even if the task seems nearly impossible, we should probably continue opposing numerical assessment wherever we can. Educational numberese may be disabling the learner rather than empowering. This text should be read as an exhortation to disable it in return.

REFERENCES

Ben-Yehuda, M., Lavy, I., Linchevski, L., & Sfard, A. (2005). Doing wrong with words: What bars students' access to arithmetical discourses. *Journal for Research in Mathematics Education, 36*(3), 176-247.

Bakhtin, M. (1986). *Speech genres and other late essays* (V. W. McGee, Trans.). Austin: University of Texas Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working inside the black box* London, UK King's College London School of Education.

Cooper, D. (2006). *Talk About Assessment: Strategies and Tools to Improve Learning*. Toronto, ON :Thomson Nelson, Government of British Columbia.

Edwards, D., & Potter, J. (1992). *Discursive psychology*. Newbury Park, CA: Sage Publications.

Halliday, M.A.K., (1987). Language and the Order of Nature. In N. Fabb, D. Attridge, A. Durant & C. McCabe (Eds.), *The Linguistics of Writing, Arguments between Language and Literature* (pp. 135-154). New York: Methuen

Harré, R., & Gillett, G. (1995). *The discursive mind*. Thousand Oaks, CA: Sage Publications.

Lave, J. (1988). *Cognition in practice*. New York: Cambridge University Press.

Sfard, A. (2007). When the rules of discourse change, but nobody tells you: Making sense of mathematics learning from a cognitive standpoint. *Journal for Learning Sciences, 16*(4), 567–615.

Sfard, A. (2008). *Thinking as communicating: Human development, the growth of discourses, and mathematizing*. Cambridge, UK: Cambridge University Press.

Sfard, A., & Prusak, A. (2005). Telling identities: In search of an analytic tool for investigating learning as a culturally shaped activity. *Educational Researcher, 34*(4), 14-22.

Vygotsky, L. S. (1987). Thinking and speech. In R. W. Rieber & A. C. Carton (Eds.), *The collected works of L. S. Vygotsky*. New York: Plenum Press.

Wang, J., & Lin, E. (2005). Comparative studies on US and Chinese mathematics learning and the Implications for Standards-Based Mathematics Teaching Reform. *Educational Researcher, 34*(5), 1-13.

Wittgenstein, L. (1953/2003). *Philosophical investigations: the German text, with a revised English translation* (G. E. M. Anscombe, Trans. 3rd ed.). Malden, MA: Blackwell Publishing.