# How common standards can diminish collective intelligence: a computational study

May 19, 2016

Forthcoming in the *Journal of Evaluation in Clinical Practice*, special issue on Philosophy of Medicine

**Michael Morreau**: michael.morreau@uit.no

**Aidan Lyon**: aidanlyon@gmail.com

**Abstract**. Good decision making depends on having accurate information—quickly, and in a form in which it can be readily communicated and acted upon. Two features of medical practice can help: deliberation in groups and the use of scores and grades in evaluation. We study the contributions of these features using a multi-agent computer simulation of groups of physicians. One might expect individual differences in members' grading standards to reduce the capacity of the group to discover the facts on which well-informed decisions depend. Observations of the simulated groups suggest on the contrary that this kind of diversity can in fact be *conducive* to epistemic performance. Sometimes, it is adopting common standards that may be expected to result in poor decisions.

**Keywords**. evidence; evaluation; computer modeling; collective intelligence

## 1 Introduction

It has long been known that collecting together inputs from several people can increase the probability of correct decisions and the accuracy of judgments (Condorcet 1785, Galton 1907). More recently, it has become clear that including different perspectives and ways of thinking within the group can be just as important as individual expertise (Surowiecki 2004, Page 2008). These insights provide an

epistemological rationale for much practice that has evolved in medicine, where knowledge is pieced together in morning reports, case conferences, peer to peer consultations, and in other meetings and discussions among staff. Improvements in the quality and efficiency of healthcare may be expected to result from increased reliance on collective intelligence in medicine (Wolf et al. 2015).

Scores and grades play a central part in medical evaluation. For example, Wells and Geneva scores are used in clinical medicine to determine pre-test probabilities for pulmonary embolism, expressed in the qualitative probability grades *low*, *intermediate*, and *high* (Wells et al. 1998, Wicki et al. 2001; decisions about further diagnostic testing depend on the outcome.[1] Following the GRADE method, systematic review panels categorize bodies of evidence as *high*, *moderate*, *low* or *very low* in quality (Balshem et al. 2011). From Apgar scoring of newborn babies (Finster and Wood 2005) to the Glasgow Coma Scale used to identify organ donors (Teasdale and Jennett 1974), scores and grades express many different aspects of health throughout our lives.

Using grades and other qualitative language brings an important advantage when decisions are made under pressure of time. It's not possible to know, say, *exactly* what the probability is that a patient who's just come into the emergency room has a pulmonary embolism. That's not necessary, though, in order to judge that the probability is *high*, or *intermediate*, or *low*, and to decide on that basis whether diagnostic procedures such as a CT scan, a chest x-ray or blood tests will be carried out. The reason is just that qualitative probabilities are *coarse grained*, each covering a range of precise probabilities. There would appear to be drawbacks to grading as well, though. One is that expressions such as *high*, *intermediate* and *low* can mean different things to different people. Large differences between doctors and their patients have been documented in the medical literature (see for instance Ohnishi et al. 2002). Similar differences were found among members of a science panel in the Netherlands (Wardekker et al. 2008), and among students of business and the social sciences (Figure 1, Morgan 2014).

One solution to problems arising from different interpretations of grades is to establish standard grading procedures for people to use. Take for instance the Geneva and Wells rules for determining pre-test probabilities. They require scoring

[1] A pre-test probability is the subjective probability that a patient has some given condition, before a diagnostic test result is available.

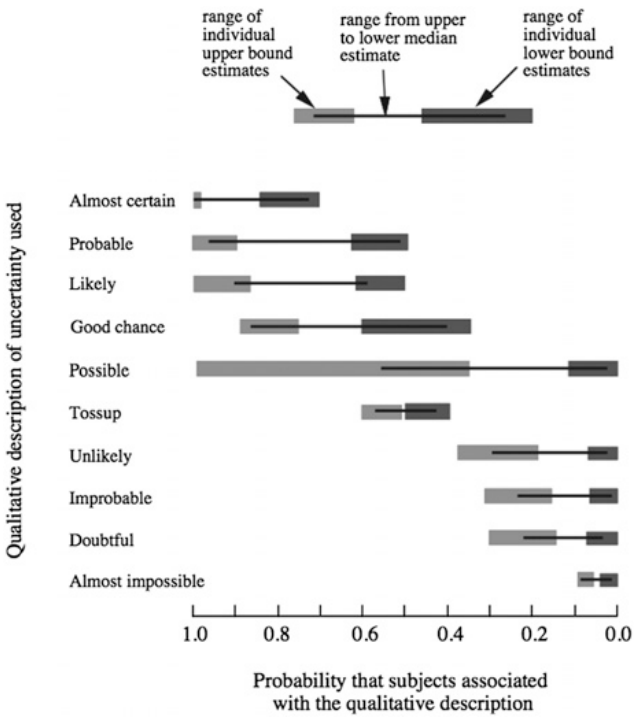


Figure 1: Reproduced from Morgan 2014.

patients on relevant criteria, and then adding up the scores to determine which grade applies (Torbicki et al. 2008, Table 7). Using the GRADE method, bodies of evidence are categorized initially on the basis of study design, and then up- or down-graded according to relevant strengths or weaknesses of the study, such as large effect size, or a serious risk of bias (Balshem et al. 2011, Table 3).

Another approach is to stipulate explicit definitions of grading expressions. In some cases it is possible to specify precise thresholds for applicability, as the Intergovernmental Panel on Climate Change (IPCC) has done for the qualitative probability grades used in its publications (Mastrandrea et al. 2011). With a multi-faceted notion such as quality of evidence it is not straightforward to specify thresholds. However, it is possible to characterize grades in ordinary language. For example, the GRADE working group specifies that *moderate* quality evidence means:

> We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different. (Balshem et al. 2011, Table 2)

Arguably these methods can be effective in establishing and maintaining common grading standards. Inter-rater reliability studies tend to confirm this (Mustafa et al. 2013, Iles et al. 2003). The potential benefits are obvious. Following standard protocols makes decisions transparent. Also, having everybody "on the same page" helps people to communicate the knowledge on which good decision-making depends.

Grading standards are conventional. In the end, it's completely up to us, say, just which precise probabilities will count as *low*, which as *intermediate*, and which as *high*. We argue, though, that they are not *merely* conventional, in that which standards we settle on can greatly affect our ability to obtain knowledge and make good decisions.[2] Indeed, we argue, some standards are actually *harmful*, in that a group that has adopted them will be worse at "tracking the truth", even, than a group whose members could have any understandings at all of what the applicable grades mean. Furthermore, we argue, which standards are better than which, and which standards are harmful, can be hard to tell. This is because it can turn on the nature of the facts that are to be discovered. For example, when these concern the pre-test probabilities of several disorders that a patient might have, it can turn on

[2]Similarly, legal speed limits are not *merely* conventional. They have consequences for public safety, levels of pollution and the efficiency of our transport system. It's different with a simple matter of coordination such as which side of the road we drive on.

what these probabilities happen in fact to be.

The capacity of a group to discover facts naturally depends on much more than just its conventions about language and the nature of these facts. The level of expertise of individual group members also matters. So does the size of the group. To help us study complex interactions between these and other factors we have built a computer model. It simulates a group with the task of ranking some given possibilities in order of their probability, on the basis of opinions expressed by individual group members using qualitative probability grades. The group's epistemic performance in this task is reckoned as the frequency with which an event that the group *judges* to be be most probable *really is* that; and by setting different parameters we can observe the consequences of the group's adopting various conventions about the meaning of the grades in which individual inputs are expressed. These observations are the basis for our claims about the epistemic merits and demerits of grading standards.

We proceed as follows. Section 2 explains our model of grading in groups and certain assumptions of its implementation. Section 3 compares the performance of simulated groups of graders working with different grading standards. Section 4 summarizes the conclusions we draw from these results.

## 2 The Model

We mentioned in the previous section the importance in medicine of scoring and grading by groups. In this section, we focus on a simple and idealized example from clinical medicine. First we build a model of this example, based on the Arrow-Sen framework in social choice theory (Arrow 1951, Sen 1970). Then we study the model with the aid of computer simulations. An important innovation is that our model includes the language of grades in which individual people express their judgments, along with their interpretations of this language. Taking interpretations of language into account is necessary for a proper understanding of collective decisions more generally, not just in the present medical context (Morreau and Weymark 2016).

First, let us introduce the example we have in mind. We imagine a set of three physicians: Brown, Jones and Smith, and a patient who presents with a set of symptoms: chest pains and shortness of breath. There is a set of possible disorders:

(H) the patient has hyperventilation, or (HA) is having a heart attack, or (P) has pneumonia, or (PE) has a pulmonary embolism, or.... The physicians are to rank the different ones, on the basis of the available information, in order of their pre-test probabilities.[3] Their goal—in this simple and idealized example—is to identify the disorder with the highest pre-test probability.[4]

We now describe a mathematical model of this simple kind of case. Let $N$ be the set of physicians, e.g.: $N = \{Brown, Jones, Smith\}$, and $D$ be the set of disorders to consider, e.g.: $D = \{d_1, d_2, ..., d_n\}$. Since the physicians will evaluate these diagnoses using a using a *grading language*, we have to specify the details of this language. There are three components to a grading language, $\mathcal{L} = (\mathcal{G}, \succ, \mathcal{I}_i)$. The first component, $\mathcal{G}$, is the set of grades that the physicians will use. For example, we might have:

$$\mathcal{G} = \{high, medium, low\}.$$

The second component of the language, $\succ$, is a linear ordering of the grades. For example, a natural linear ordering of the above grades would be: $high \succ medium \succ low$.

The third component $\mathcal{I}_n$ is the one we have most to say about here. $\mathcal{I}_n$ is an *interpretation function* for individual physician $n \in N$ of the grades in $\mathcal{G}$. It assigns to each $g \in \mathcal{G}$ this individual's interpretation of $g$. For example, $\mathcal{I}_{Brown}$ might be the function:

$$\mathcal{I}_{Brown}(high) = [66\%, 100\%]$$
$$\mathcal{I}_{Brown}(medium) = [33\%, 66\%)$$
$$\mathcal{I}_{Brown}(low) = [0\%, 33\%)$$

Intuitively, $\mathcal{I}_n$ tells us how individual $n$ understands the different available grades. In the above example, physician Brown understands the probability grade *high* to

[3]The pre-test probability $P(D^+)$ of a disorder is understood here to be the proportion $D^+/(D^+ + D^-)$ of patients with this disorder among all those with this particular patient's symptoms.

[4]We assume that identifying the most probable condition is relevant to medical decision making. Sometimes it certainly is: in determining the Wells score for PE, for instance, physicians are expected to determine whether PE is the most probable diagnosis—or, anyway, whether it is at least as likely as any other (Wells et al. 1998). In general perhaps it is more important instead to identify those possibilities that require immediate action; for now, though, we assume that what matters is simply probability.

correspond to the range of pre-test probabilities from 66% to 100%. Of course, different individuals may have different interpretations of the grades. For example, Jones might interpret the grades as follows:

$$\mathcal{I}_{Jones}(high) = [90\%, 100\%]$$
$$\mathcal{I}_{Jones}(medium) = [10\%, 90\%)$$
$$\mathcal{I}_{Jones}(low) = [0\%, 10\%)$$

If $\mathcal{I}_m = \mathcal{I}_n = \mathcal{I}$, for every $m, n \in N$, then $\mathcal{I}$ is called a *consensus interpretation*. Intuitively, if there is a consensus interpretation, then everyone in the group has the same understanding of what the grades mean. Such a consensus may arise naturally or it may arise because the physicians have been given explicit instruction to use a particular interpretation.

In real situations, physicians often (but not always) use scoring systems, such as the Wells and Geneva systems, to assign grades to possible diagnoses. We could model these systems explicitly. Instead we take a more general approach that abstracts away from the details of whatever scoring system or other process it is by which physicians assign probability grades. We think of it like this: physicians are presented with some information, $i$. It includes the patient's symptoms, medical history and so on. For each disorder $d$, they convert this information $i$ into a corresponding grade $g$. Mathematically, we can think of this process as the following mapping:

$$(i, d) \longrightarrow g$$

This mapping might be instantiated via a scoring system that the physician has been instructed to use or it might be instantiated by some other process — e.g., the physician might make an intuitive judgment of $i$ and $d$ to assign grade $g$.

In order to have a model that we can study systematically, we now add some details. First, we assume that $i$ and $d$ together fix some objectively correct pre-test probability $p \in [0, 1]$. Second, we introduce the notion of an *individual point estimate*. The idea is that for any given individual $n$ there is some noisy process that turns $p$ into $n$'s estimate $e$ thereof. For simplicity, we assume this process is an unbiased Gaussian with variance $\sigma^2$ (i.e. a normal distribution or "bell curve"). Third, we assume that $e$ is then converted to the appropriate grade $g$, according to $n$'s interpretation $\mathcal{I}_n$. For example, if $e = 5\%$ and we use Jones' above interpretation,

then the grade Jones assigns to the disorder in question is *low*. The model, then, can be thought of as the following sequence of mappings:

$$(i, d) \longrightarrow p \longrightarrow e \longrightarrow g$$

Note that we do not assume that $n$ is *aware* of $e$ or actually constructs $e$ in any conscious way. Indeed, we think that in many situations this is rather unlikely (scoring systems typically do not mention point probabilities). Instead, we take the above to be an abstract model of individual grading. As such, $e$ is a *theoretical* parameter that we use to generalize across many different ways of moving from $(i, d)$ to $g$, including the use of scoring rules and other protocols that make no reference to point probabilities or estimates.

Based on this noisy process and the individual interpretations of the grades, the individuals assign grades to each of the disorders. Each individual then selects the disorder they give the highest grade. (We must allow for the possibility that several disorders get the top grade. For example, Brown might think that both pneumonia and pulmonary embolism are *high* probabilities. In such situations, we assume that ties are broken by randomly selecting from the top grade category.) Either the individual's selection is correct, being at the top of the objective ordering, or else it is not. The individual's *expected performance* is, by definition, the proportion of cases in which their selection is correct, within a large number of trials.

So far we have been busy with individual judgments. However, we are interested the judgments of the *group*. These are gotten by *aggregating* the individually assigned grades. Different procedures can be used for this. Here we follow the simple method of taking the *median* or middlemost of the individually assigned grades: the grades assigned to $d$ by the different $i \in N$ are first listed in order of $\succ$ (including repetitions, when several have assigned the same grade). Assuming the number of $N$ is odd, the collective grade for $d$ is the one that's in the middle of the list. If $N$ is even, the collective grade for $d$ is randomly selected from the two grades in the middle of the list.[5]

The group chooses the disorder to which it gives the highest grade. (Again, ties are broken by random selection.) Either the group's selection is correct, in that this disorder really is one of those with the highest pre-test probability, or else it is not.

[5]One advantage of making decisions in groups is that people's positions change in light of group discussion. By simply aggregating the individually assigned grades we leave this important aspect of deliberation to one side. Modeling it is a promising line for future work of the sort begun here.

The group's *expected performance* is, by definition, the proportion of cases in which its selection is correct, within a large number of trials.

In this paper we focus on individual interpretations of the probability expressions. Would it be good for the group to have a single common interpretation—a semantic *consensus*, as we will call it? If so, which semantic consensus will be best, in the sense that it is most favorable for determining pre-test probabilities? Even in our simple example, this depends in complicated ways on multiple factors. In order to answer such questions as these we take up a tool that is often used to study complex social interactions: multi-agent computer simulation (Epstein 2007). That requires making further modeling assumptions, to which we now turn.

For one thing, we need an assumption about how the pre-test probabilities of the different possible disorders are distributed. We don't assume the disorders are mutually exclusive: a patient can have several of them at once. For simplicity, and for some degree of realism, we assume that the pre-test probabilities are distributed according to a Pareto distribution, $\alpha x^{\alpha}/x^{\alpha+1}$, where $\alpha$ is a parameter that we can vary in the simulations.[6] For each simulation run of our model, a distribution of 30 pre-test probabilities is generated from a Pareto sample, the individuals then grade the diagnoses according to the model described above with $\sigma^2 = 10\%$, and then their performance is assessed in terms of how often they choose disorders with the highest pre-test probability. Finally, the group grades are determined by aggregating the individual grades and the group performance is then also assessed in terms of how often it chooses the most probable disorders. Performance scores are averaged over 2,000 simulation runs.

Regarding the grades and interpretations, we assume that the physicians have 3 grades available to them (this is realistic: many real scoring and grading protocols use from 2 to 4). And in our simulations, we focus on the three kinds of consensus interpretations and one non-consensus set of interpretations:

1. A *symmetric* consensus interpretation with thresholds of $[33, 66, 100]$. (That is, individuals with this consensus interpret *low* as $[0, 33]$, *medium* as $[33, 66]$, and *high* as $[66, 100]$.)

2. A *bottom-heavy* consensus interpretation with thresholds of $[25, 50, 100]$.

[6] With a Pareto distribution, the different diagnoses can be "bunched up" at the bottom, with many improbable ones and only a few that are likely. See for example figure 2. Smaller values of $\alpha$ make more diagnoses more likely.

3. A *top-heavy* consensus interpretation with thresholds of $[50, 75, 100]$.

4. A random set of interpretations, where each individual has thresholds $[a, b, 100]$, where $a$ and $b$ are generated by randomly selecting uniformly from the $(0, 100)$ interval.

Our goal is not to study every possible interpretation, but rather to demonstrate how the performance associated with a particular consensus can depend dramatically on the details of the situation at hand. Similarly, although we've made many precise and idealising assumptions — e.g., that the pre-test probabilities are approximately Pareto — our goal is to show how the performance associated with a particular consensus is highly context sensitive. The upshot then is that when we are designing scoring and grade systems, we need to think carefully about what kinds of situations they are going to employed in, and that computer simulations can help us with this task.

## 3 Results

We studied 2,000 simulation runs of the model described in the previous section for two parameter settings of the Pareto distribution from which the pre-test probabilities are sampled: $\alpha = 3$ and $\alpha = 20$. These are summarised in figures 2 and 3. Recall that smaller values of $\alpha$ make more disorders more likely. So we have effectively studied two kinds of situations: one in which the vast majority of disorders are unlikely ($\alpha = 20$) and one in which several disorders tend to have a pre-test probability greater than 50% and some are even quite likely, with probabilities above 80% ($\alpha = 3$).

In the first kind of situation (figure 2), the bottom-heavy consensus — i.e., the "[25,50,100]" consensus — gives rise to the best performance (for both individuals and groups) and the top-heavy consensus — i.e., the "[50,75,100]" consensus — gives rise to the worst performance (again, for both individuals and groups). However, in the second kind of situation (figure 3), the exact reverse occurs: the top-heavy consensus gives rise to the best performance (for both individuals and groups) and the bottom-heavy consensus gives rise to the worst performance (again, for both individuals and groups). This reason is as follows. In the first kind of situation, the bottom-heavy consensus makes more distinctions where most of the

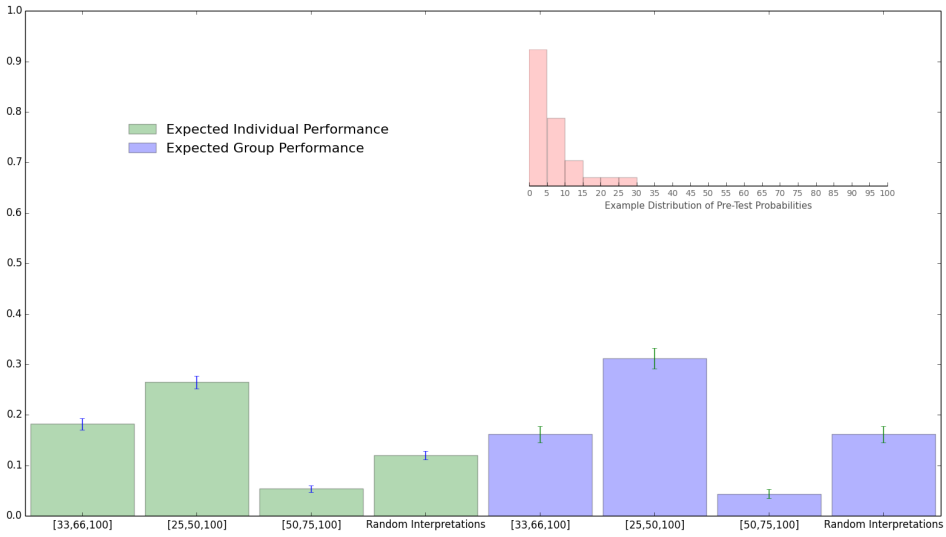


Figure 2: Performances of different interpretations for individuals (green) and groups (blue) for a Pareto distribution of pre-test probabilities with $a = 20$. Error bars are 95% confidence intervals based on 2,000 simulation runs.

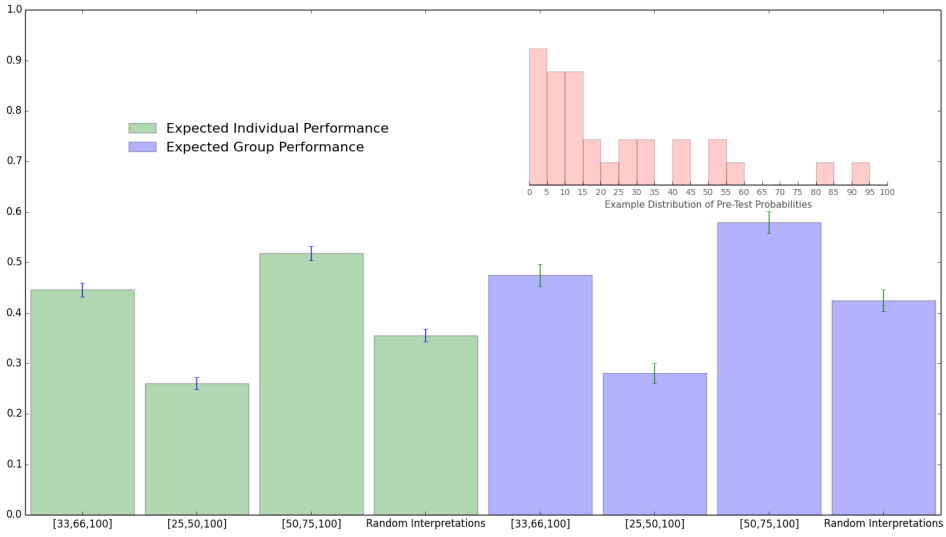


Figure 3: Performances of different interpretations for individuals (green) and groups (blue) for a Pareto distribution of pre-test probabilities with $a = 3$. Error bars are 95% confidence intervals based on 2,000 simulation runs.

pre-test probabilities are — see the example distribution of the pre-test probabilities in the figure 2). And, in the second situation, the top-heavy consensus makes more distinctions where they matter: where the very likely pre-test probabilities are — see the example distribution of the pre-test probabilities in the figure 3).

This reversal demonstrates just how sensitive the performance associated with a particular consensus can be. The only difference between the two kinds of situations is the variation of the parameter $\alpha$ from 20 to 3. In real life, situations will vary in many more — and much more complicated — ways.

In both situations, some semantic consensus resulted in worse epistemic performance than random interpretations: in figure 2, the top-heavy consensus was worse; in figure 3, the bottom-heavy consensus was. This suggests that if time and energy are to be spent on forming a semantic consensus, it is important to think carefully about which it will be. Otherwise, it might well be better simply to let everyone interpret the grades as they please. Surprisingly, perhaps, letting people understand the applicable grades any old how can result in better performance than reaching a common understanding.

In both situations we see collective wisdom effects — that is, group performance tended to be better than individual performance — but they are small. This is mostly because the group of physicians was assumed to be small, with just three members, and because the noise associated with the process of generating the hypothetical estimates was fairly low ($\sigma^2 = 10\%$). Clearer differences between individual and group performance can be observed with more noise (e.g., $\sigma^2 = 30\%$) and a larger group (e.g., $|N| = 10$). This is demonstrated in figure 4.

In the cases we have looked at so far, there is always a consensus interpretation that performs better than the symmetric consensus. However, by changing some of the assumptions of our model we find situations in which the symmetric consensus performs best (of the interpretations we discuss here). For example, changing the underlying distribution of pre-test probabilities to a Gaussian, with a mean of 50% and a variance of 10%, and reducing the number of possible disorders from 30 to 10 (figure 5), we observe that the symmetric consensus performs better than the other consensus interpretations and random interpretations. Again, we see how the relative performance of a consensus interpretation is sensitive to the details of the situation at hand.

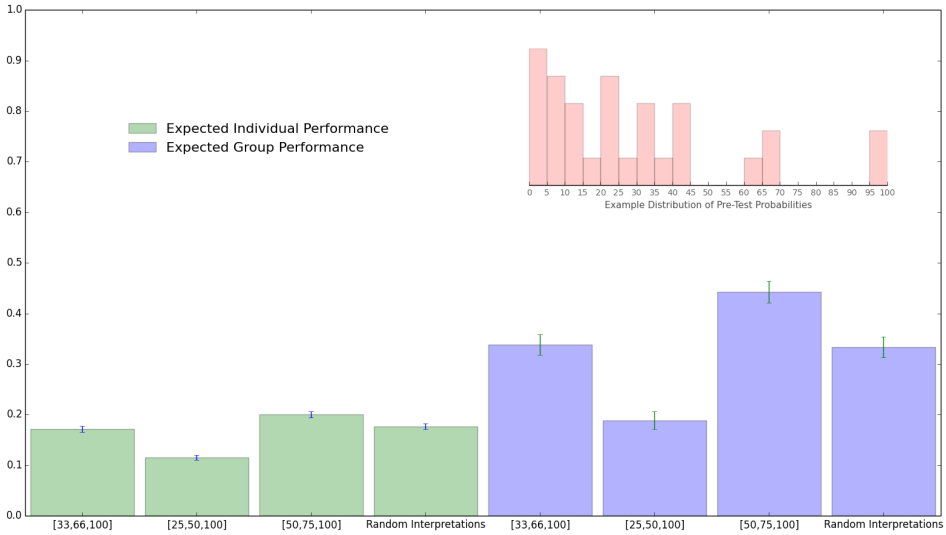


Figure 4: Performances of different interpretations for individuals (green) and groups (blue) for a Pareto distribution of pre-test probabilities with $a = 3$, noisier physicians ($\sigma^2 = 30\%$) and more physicians ($|N| = 10$). Error bars are 95% confidence intervals based on 2,000 simulation runs.

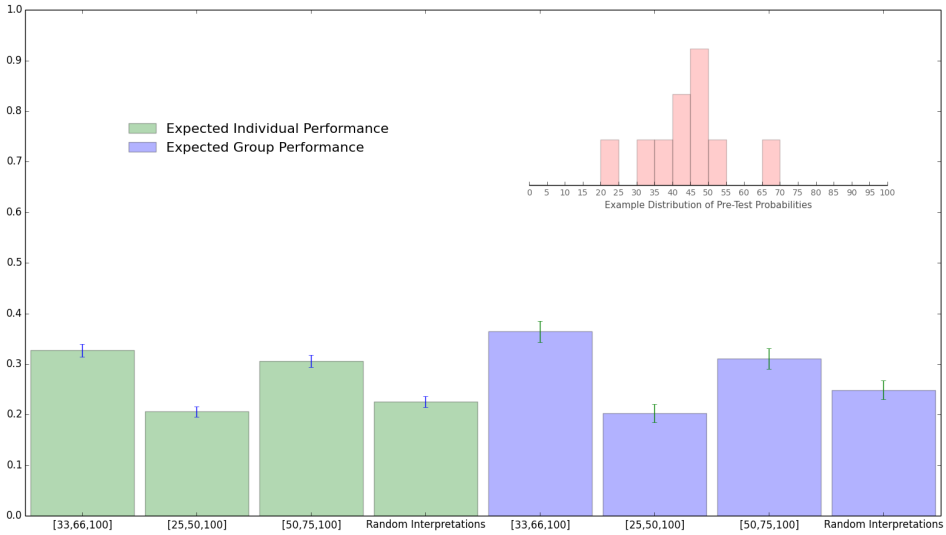


Figure 5: Performances of different interpretations for individuals (green) and groups (blue) for a Gaussian distribution with mean 50% and variance 10% and 10 possible disorders being evaluated by the physicians. Error bars are 95% confidence intervals based on 2,000 simulation runs.

## 4 Summary and discussion

People often find it easier to grade, and to communicate with others in qualitative terms, rather than to use more rigorous language. In order to avoid interpersonal differences in the interpretation of grades, steps can be taken to standardize usage. These include the use of scoring and grading protocols, and providing explicit definitions of the relevant expressions.

We showed using a computer simulation that, from an epistemological viewpoint, some standards are better than others in some contexts and worse in others. Simulated individuals and groups with the epistemically better standards are better at discovering relevant facts. Importantly, some standards can result in rather poor performance, in that a group having adapted them may be expected to be less capable of discovering facts, even, than a group of people whose members could interpret the grades in any way at all. (This can be seen in figures 2 and 3, where the performances of random interpretations were better than top-heavy and bottom-heavy interpretations, respectively.) Furthermore, and crucially, which standards lead to better performance than which, and whether adopting any given standards would help the group or hurt it, is highly contextual. In the case we considered here, it depends, among many other things, on what the probabilities of the given disorders happen to be.

Our model of individual and collective grading is simple. Surely some of its assumptions and idealizations will be found unrealistic. The task we analyzed, choosing the most-probable disorder from among several, is also simple. A more important task, perhaps, is choosing the disorder with the highest expected danger, since although such disorders might be relatively unlikely they can require immediate action given their high-level of danger. Since some disorders with substantial danger are more likely than others, this should have a significant impact on the performances associated with different grading standards. With multi-dimensional tasks there is also further scope for interpersonal differences and for standardization, since in addition to having different thresholds people can attach different weights and priorities to the several dimensions. There is much work to be done before our simulation studies cover any large part of real medical practice.

Be this as it may, the results already point to the contribution that computational studies can make. Scores, grades and other qualitative language are used throughout medicine, for estimating probabilities, evaluating evidence and more.

Our simulations show not only that different ways of standardizing such language can have very different consequences for performance, but also that many other factors are relevant, all interacting in complicated and sometimes unintuitive ways. Introducing computational studies and statistical analysis into the development of scoring and grading systems may be expected to help researchers to think through the consequences of basic design decisions such as how many grades to use, how many people are needed for expert panels to be effective and what the required levels of expertise are—*before* their decisions become established in medical practice.

## References


Arrow, K. J. (1951). *Social Choice and Individual Values*. New York: Wiley. Second edition 1963.

Balshem, H., M. Helfand, H. J. Schnemann, A. D. Oxman, R. Kunz, J. Brozek, G. E. Vist, Y. Falck-Ytter, J. Meerpohl, S. Norris, and G. H. Guyatt (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology 64*(4), 401–406.

Condorcet, J.-A.-N. d. C. (1785). *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix [microform] / par M. le Marquis de Condorcet*. Imprimerie royale Paris.

Epstein, J. M. (2007). *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press.

Finster, M. and M. Wood (2005, April). The Apgar score has survived the test of time. *Anesthesiology 102*(4), 855857.

Galton, F. (1907). Vox Populi. *Nature 75*, 450–1.

Iles, S., A. Hodges, J. Darley, C. Frampton, M. Epton, L. Beckert, and G. Town (2003). Clinical experience and pre-test probability scores in the diagnosis of pulmonary embolism. *QJM 96*(3), 211–215.

Mastrandrea, M., K. Mach, G.-K. Plattner, O. Edenhofer, T. Stocker, C. Field, K. Ebi, and P. Matschoss (2011). The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups. *Climatic Change 108*(4), 675–691.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proc. Natl. Acad. Sci. USA 111*(20), 7176–7184.

Morreau, M. and J. A. Weymark (2016). Measurement scales and welfarist social choice. *Journal of Mathematical Psychology*. In press.

Mustafa, R. A., N. Santesso, J. Brozek, E. A. Akl, S. D. Walter, G. Norman, M. Kulasegaram, R. Christensen, G. H. Guyatt, Y. Falck-Ytter, S. Chang, M. H. Murad, G. E. Vist, T. Lasserson, G. Gartlehner, V. Shukla, X. Sun, C. Whittington, P. N. Post, E. Lang, K. Thaler, I. Kunnamo, H. Alenius, J. J. Meerpohl, A. C. Alba, I. F. Nevis, S. Gentles, M.-C. Ethier, A. Carrasco-Labra, R. Khatib, G. Nesrallah, J. Kroft, A. Selk, R. Brignardello-Petersen, and H. J. Schnemann (2013). The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. *Journal of Clinical Epidemiology 66*(7), 736 – 742.e5.

Ohnishi, M., T. Fukui, K. Matsui, K. Hira, M. Shinozuka, H. Ezaki, J. Otaki, W. Kurokawa, H. Imura, H. Koyama, and T. Shimbo (2002). Interpretation of and preference for probability expressions among Japanese patients and physicians. *Family Practice 19*(1), 7–11.

Page, S. (2008). *The Difference*. Princeton University Press.

Sen, A. K. (1970). *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

Surowiecki, J. (2004). *The Wisdom of Crowds*. Doubleday.

Teasdale, G. and B. Jennett (1974). Assessment of coma and impaired consciousness: a practical scale. *The Lancet 304*(7872), 81 – 84. Originally published as Volume 2, Issue 7872.

Torbicki, A., A. Perrier, S. Konstantinides, G. Agnelli, N. Galiè, P. Pruszczyk, F. Bengel, A. J. Brady, D. Ferreira, U. Janssens, W. Klepetko, E. Mayer, M. Remy-Jardin, J.-P. Bassand, A. Vahanian, J. Camm, R. De Caterina, V. Dean, K. Dickstein, G. Filippatos, C. Funck-Brentano, I. Hellemans, S. D. Kristensen, K. McGregor, U. Sechtem, S. Silber, M. Tendera, P. Widimsky, J. L. Zamorano, J.-L. Zamorano, F. Andreotti, M. Ascherman, G. Athanassopoulos, J. De Sutter, D. Fitzmaurice, T. Forster, M. Heras, G. Jondeau, K. Kjeldsen, J. Knuuti, I. Lang, M. Lenzen, J. Lopez-Sendon, P. Nihoyannopoulos, L. Perez Isla, U. Schwehr, L. Torraca, and

J.-L. Vachiery (2008). Guidelines on the diagnosis and management of acute pulmonary embolism. *European Heart Journal 29*(18), 2276–2315.

Wardekker, J. A., J. P. van der Sluijs, P. H. M. Janssen, P. Kloprogge, and A. C. Petersen (2008). Uncertainty communication in environmental assessments: views from the Dutch science-policy interface. *Environmental Science and Policy 11*, 627–641.

Wells, P. S., J. S. Ginsberg, D. R. Anderson, C. Kearon, M. Gent, A. G. Turpie, J. Bormanis, J. Weitz, M. Chamberlain, D. Bowie, D. Barnes, and J. Hirsh (1998). Use of a clinical model for safe management of patients with suspected pulmonary embolism. *Annals of Internal Medicine 129*(12), 997–1005.

Wicki, J., T. Perneger, A. Junod, H. Bounameaux, and A. Perrier (2001). Assessing clinical probability of pulmonary embolism in the emergency ward: A simple score. *Archives of Internal Medicine 161*(1), 92–97.

Wolf, M., J. Krause, P. A. Carney, A. Bogart, and R. H. Kurvers (2015). Collective intelligence meets medical decision-making: The collective outperforms the best radiologist. *PloS one 10*(8), e0134269.