Supergrading: how diverse standards can improve collective performance in ranking tasks

Michael Morreau

Department of Philosophy, UiT – The Arctic University of Norway, Postboks 6050, Langnes, 9037 Tromsø, Norway

Abstract

The method of supergrading is introduced for deriving a ranking of items from scores or grades awarded by several people. Individual inputs may come in different languages of grades. Diversity in grading standards is an advantage, enabling rankings derived by this method to separate more items from one another. A framework is introduced for studying grading on the basis of observations. Measures of accuracy, reliability and discrimination are developed within this framework. Ability in grading is characterized for individuals and groups as the capacity to grade reliably, accurately and at a high level of discrimination. It is shown that the collective ability of a supergrading group with diverse standards can be greater than that of a less diverse group whose members have greater ability.

Keywords: Scoring and Grading, Social Choice, Cognitive Diversity, Collective Intelligence

Many decisions call for items of one kind or another to be ranked along some relevant dimension. Project proposals are ranked in order of funding priority, possible outcomes by their probability, hotels and restaurants by how good they are, and so on. One way to rank items is by grading them. The grades could be numerical scores, qualitative probability terms such as *likely*, *tossup* and *unlikely*, numbers of stars or any other expressions ordered from top to bottom. Items are ranked by the top-bottom order of their grades; those with the same grade are tied.

There is good reason to rank things by grading them. People often find it difficult to arrive at precise judgments. Information sometimes is missing, and always time is short. In many instances they can even so assign scores or grades, though, because these are coarse grained; you don't need to pinpoint the probability of an outcome to be satisfied that it is *likely*, for instance, because this term covers a range of precise probabilities. Grading helps people to contribute judgments that, however imprecise, are more-or-less *accurate*. The assigned grades are the right ones, or close.

The accuracy does not come for free. The less precise the grading scale, the more ties there are; there is a cost in *discrimination*, the separation of different items by different grades. How much that matters depends on the case: a finer ranking might be needed to pick out the very best project proposal or job candidate, perhaps, than for funding or shortlisting some of the better ones. One basic question is therefore how to manage the accuracy-precision trade-off. How can we have graded judgments that are both sufficiently accurate *and* sufficiently precise for whatever task is at hand?

Commonly decisions are made by committees, panels and other groups. In a funding decision, for instance, it is typical for several reviewers to score the proposals individually; priorities are then based on the scores they assign. Now, putting together the judgments of several people can improve the terms of the accuracy-precision trade-off. Francis Galton in a classic study of a competition at a county fair found that the median of many precise estimates of the weight of an ox was more accurate than, on average, the individual estimates [9]. Judgments expressed as scores and grades can be aggregated in the same way. For each item, you line up all assigned grades in the top-bottom order of the grades, and then choose from the middle. This might be expected to promote accuracy just as it does with precise estimates.

There is a catch. The judgment that an ox weighs $1,234.5$ kilos means the same coming from one person as another, but people have different standards for awarding scores and grades. A score of 4 from one reviewer on a funding panel might be equivalent not to a 4 but rather a 3 from another reviewer—or else a 5 or even a 6. What order should everybody's inputs be lined up in when there are such differences to reckon with? Differences in their standards might seem to make nonsense of the idea of aggregating scores and grades awarded by different people.

The matter is not academic. Large differences in the interpretation of probability grades have been documented even among people who are culturally and linguistically quite similar. These include students [21], doctors and their patients [18], and members of science panels and boards [22, 16, 17].

There are ways to cope. Sometimes it is possible to specify thresholds in completely precise terms, as the *Intergovernmental Panel on Climate Change* has done for qualitative probabilities in its publications [15].[1] Grading protocols constrain understandings to some extent through guidance in proper use [3]. Individual judgments can be corrected for interpersonal differences using techniques from polling such as anchoring vignettes [14]. Where common understandings are found or manufactured, median grading will under favourable circumstances boost accuracy. So will majority judgment, a refinement of median grading with an innovative method for breaking ties [1, 2].

Be this as it may, diverse understandings of scores and grades don't only make trouble. They also create opportunities. This article introduces *supergrading*, a new method for aggregating individually assigned grades. The resulting *supergrades* are no more accurate than the individual inputs. But in general they are more precise and informative, resulting in rankings that separate more items from one another. Diverse grading standards are an asset with this method because they are what boost discrimination. An example illustrates.

1. An ox-ranking task

With a nod to Galton, imagine another competition at a fair. There is a herd of oxen and whoever picks out the heaviest wins. Now, you and I are not farmers. We are not butchers or livestock auctioneers, and cannot hope to pinpoint the weights of the different ones. The best we can do is to say imprecise things such as "this one looks *heavy* to me", and "that other one I'd say is *light*, for an ox". Do people like us even stand a chance?

Suppose your judgments, however imprecise, are accurate *by your standards*: when you say an ox is *heavy* its weight always falls within the range that, as you understand it, is covered by this expression; and if you say *light* its weight is always compatible with your understanding of that. My own understanding of these expressions is different. You and I draw the line between *heavy* and *light* at different places. But, suppose, my judgments too are accurate, taken on their own terms.

Then you and I together can divide the oxen into three categories by weight, though each of us uses just the two expressions to describe them. Say for concreteness that you draw the line between *heavy* and *light* at 1000 kilos, and I at just 500 kilos. When both of us call an ox *heavy* it must weigh

[1] Specifying precise thresholds cannot be counted on entirely to remove interpersonal differences ([5], [6]).

at least 1000 kilos. If on the other hand we disagree then the weight must be at least 500 but under 1000 kilos, and if we agree that an ox is *light* then it weighs under 500 kilos. More of us can do even better: three binary graders with pairwise differing standards can discriminate four categories, and so on.

So it is that people with little expertise do indeed stand a chance. We can rank the oxen in the correct order and pick the heaviest one—provided there are enough of us, and our interpretations of language are diverse.

Think of us as expressing ourselves, collectively, in a richer language than either of us uses separately. When you say an ox is *light* and I say *heavy*, you and I together, by that fact, count it *light-heavy*. We assign a *supergrade*, the concatenation of individually assigned grades. Our other supergrades are *heavy-heavy*, at the top, and the lowest grade *light-light*.

This article develops the method of supergrading and explores some of its consequences for collective decision making. Some noteworthy features are already visible. First, supergrades are more precise than the individually assigned grades that make them up. Our supergrades distinguish for example among your *light* oxen (up to 1000 kilos) the heavier ones that are *light-heavy* (from 500 up to 1000 kilos) and lighter ones that are *light-light* (up to 500 kilos).

Second, there's no need for a common language of grades. You and I used the same terms *heavy* and *light* but we needn't have done. *Any* scores or grades whatsoever make up supergrades, no matter how diverse they are in number or interpretation, provided only that they measure the same dimension, here weight. You could just as well have scored the oxen from 1 to 10, say, or I could have graded them as *large*, *medium* or *small*. This tolerance with respect to the form of inputs is conducive to tapping diverse sources of information and collective intelligence [20, 19].

Third, in order to find out the group's ranking there's no need to know just which grades people might use, or just where they draw the lines between them. What matters is each grader's top-bottom order among the grades they do in fact use: whether, say, you take that 3 you assigned to be the higher score, or the 5. The top-bottom order of the assigned supergrades is determined by this.

Finally, supergrading is not a *competitor* to accuracy-enhancing aggregation methods such as median grading. It is a precision-enhancing *complement* to them. You and I might as well be two halves of a group: everyone in one half has your understanding of the grades, and everyone in the other shares mine. Now each half aggregates its own members' inputs by taking medians; under favorable circumstances, each half grades accurately, by its own standards—even if most members do not. The halves then contribute accurate grades that make up supergrades of the group as a whole, with differences of understanding between them boosting precision.

Thus, with judicious use of both kinds of method, sufficiently large and diverse groups can reach judgments that are both accurate and precise enough for whatever task is at hand. The accuracy-precision trade-off for individuals is circumvented.

Picking the heaviest ox out of a herd by grading them all is of course just a toy example, chosen to honor Galton. In a real problem of this sort we'd put each ox on a livestock scale and be done with it. There are many real classification problems though in which the use of grades cannot be avoided, whether that is because precise cardinal information is difficult to obtain or because ordinal information is all that can be had, even in principle.

Take the measurement of severity of illness in clinical trials, held to assess the efficacy of medical interventions such as drugs. For instance, human "readers" of endoscopy videos score patients' ulcerative colitis disease activity on the endoscopy component (*normal-mild-moderate-severe*) of the Mayo Clinic Scale. Accuracy is critical because the more accurate the scores are, the better researchers can tell effective interventions from ineffective ones.

One way to increase reading accuracy that has recently been proposed is to aggregate scores assigned by several independent readers [10]. The authors propose a "2+1" collective scoring procedure that outputs the common score of two initial readers whenever they agree, and the median of their different scores together with the score of a third reader whenever they do not. Supergrading binary severity judgments is another way of achieving accuracy that could be explored. Instead of expecting individual readers to provide inputs using all four endoscopy scores of the Mayo Scale, the collective reading task could be set up so that different readers specialize on different parts of the scale. Let's say that one of three readers is responsible for determining just whether the score for a given video is above *normal*, another whether it is above *mild*, and the third whether the score is above *moderate*. Individual readers might be expected very often to achieve accuracy in this relatively undemanding task, and when their binary judgments are combined by supergrading, accurate scores on the endoscopy component of the Mayo Scale will be the result.[2]

Besides the measurement of severity of illness there are many other matters that could be approached using the method of supergrading. These include college admissions, hiring decisions, stock evaluations, the evaluation of potential markets by venture capitalists, qualitative risk assessments by engineers, ranking sports teams, classification of loans by loan officers, and more.[3] Here, the ox-grading problem serves as a running example throughout. It stands in as a model for them all.

The article develops as follows. Section 2 introduces grading languages. Section 3 characterizes grading problems and their solutions. Section 4 shows how to combine grading languages into more-precise superlanguages, and solutions into supersolutions. Section 5 states conditions under which grading problems have solutions that are both reliable and accurate. Section 6 characterizes ability in grading as reliability and accuracy together with discrimination. Section 7 shows that some groups with lower individual ability but more diversity have greater collective ability than other groups with greater individual ability but less diversity. Finally, section 8 briefly remarks on some consequences for the design of committees and expert panels, and mentions directions for future research.

2. Languages of grades

A first step is to distinguish between the signs we use for grading and our interpretations of these. These signs, or *grade terms*, come with an ordering from "top" to "bottom". Technically, a *grade vocabulary* is a pair $\langle T, \succeq \rangle$, where T is a finite, non-empty set, the grade terms, and $\succeq$ is a linear ordering of T: antisymmetric[4], transitive[5] and total.[6]

Grade terms become meaningful expressions, grades, when interpreted as intervals of some dimension of interest. Let V be a non-empty set, the *values*, with its own linear ordering $\geq$. The values can be precise weights, probabilities, degrees of merit or what have you. Let a function I determine

[2] Whether this approach results in a greater probability of accurate results than alternatives is an empirical matter that will not be addressed here.

[3] For a wealth of other practical examples, see [8].

[4] $\forall e, f \in T$, if $e \succeq f$ and $f \succeq e$, then $e = f$.

[5] $\forall e, f, g \in T$, if $e \succeq f$ and $f \succeq g$, then $e \succeq g$.

[6] $\forall e, f \in T$, either $e \succeq f$ or $f \succeq e$.

for each $e \in T$ some $I(e) \subseteq V$. I is an *interpretation* of $\langle T, \succeq \rangle$ *in* $\langle V, \geq \rangle$ if, first, for each $e \in T$, $I(e)$ is a convex set;[7] second, I partitions V;[8] and, finally, I is *orderly* in the sense that higher grades go with higher values.[9]

Consider any interpretation I of $\langle T, \succeq \rangle$ in $\langle V, \geq \rangle$. Because I partitions V there is, for any given $v \in V$, some unique corresponding term $e \in T$ such that $v \in I(e)$. We write it $I^{-1}(v)$, so $I^{-1}(v) = e$ is equivalent to $v \in I(e)$. Observe that for any $v \in V$, $v \in I(I^{-1}(v))$.

A *grade language* $L = \langle T, \succeq, I \rangle$ *for* $\langle V, \geq \rangle$ is a vocabulary $\langle T, \succeq \rangle$ together with an interpretation I of this vocabulary in $\langle V, \geq \rangle$. A language *measures* the dimension that it is for.

Let $L = \langle T, \succeq, I \rangle$ measure $\langle V, \geq \rangle$. A value $s \in V$ is a *standard* for e *in* L if $e \in T$, $s \in I(e)$, $I^{-1}(v) \succeq e$ for any $v \in V$ such that $v \geq s$, and $I^{-1}(v) \prec e$ if $v < s$. Intuitively, anything that meets the standard for a grade gets that grade or higher, and anything falling short gets a lower grade. Notice that for any given e there is at most a single standard in L. S is a *set of standards* for L if for each $s \in S$ there is some e such that s is the standard for e in L.

> *Example:* Our common vocabulary in the competition has as its terms $T = \{(h)eavy, (l)ight\}$, with h the top grade and l the bottom one: $h \succeq l$. We interpret h and l as intervals of the positive real numbers including 0. Your interpretation is: $I_{you}(h) = [1000, \infty)$ (the numbers at least 1000) and $I_{you}(l) = [0, 1000)$ (at least 0 and less than 1000). Mine is: $I_{me}(h) = [500, \infty)$ and $I_{me}(l) = [0, 500)$. We grade in distinct languages $L_{you} = \langle T, \succeq, I_{you} \rangle$ and $L_{me} = \langle T, \succeq, I_{me} \rangle$. The standard for h in L_{you} is 1000 kilos. The standard for h in L_{me} is 500 kilos.

Notice that not all grading languages have standards, in the special sense of this article. For instance, languages whose labels are interpreted as intervals with upper bounds, but no lower bounds, do not have standards. Those languages that do have standards in this special sense are a focus of

[7] $I(e)$ is *convex* if $\forall u, v, w \in V$, if $u, w \in I(e)$ and $u \geq v \geq w$, then $v \in I(e)$.

[8] I *partitions* V if $\forall e \in T, I(e) \neq \emptyset$; $\forall e, f \in T$, if $I(e) \cap I(f) \neq \emptyset$ then $e = f$; and $\bigcup\{I(e) : e \in T\} = V$.

[9] Technically, I is *orderly* if $\forall e, f \in T$ such that $e \succ f$, $I(e) > I(f)$. Here, $\succ$ is the *asymmetric component* of $\succeq$. That is, $e \succ f$ if $e \succeq f$ and $f \not\succeq e$. The asymmetric component $>$ of $\geq$ has been extended from individual elements of V to sets $S, T \subseteq V$: $S > T$ if $\forall s \in S$ and $\forall t \in T, s > t$.

attention here just because it is easy to illustrate using them a main point of this article: that interpersonal differences in the interpretation of grades can contribute to collective wisdom, by giving groups a greater capacity to discriminate among items than their individual members have. Once the reason for this has been understood, it is not difficult to see that it holds also for interpretations that do not provide languages with standards.

3. Grading problems and their solutions

Some attributes of things are *gradable*. They take values in a structure $\langle V, \geq \rangle$ of which, as before, V is a non-empty set, the values, and $\geq$ is an ordering of V. For instance, the attribute *weight* of oxen takes values in $\langle \mathbb{R}^+, \geq \rangle$, the positive real numbers, ordered in the usual way. Notice that this definition accommodates not only attributes that are cardinally measurable, such as *weight*, but also those that are merely ordinally measurable, such as *creativity* in scientific project proposals and the *friendliness* of candidates for a job.

Let X be any non-empty set, the *items*. They are oxen, project proposals, job candidates or what have you. A *grading problem* $\langle X, \alpha \rangle$ pairs X with some gradable attribute α such that, where the values of α are in $\langle V, \geq \rangle$, $\forall x \in X$, $\alpha(x) \in V$. Intuitively, α is an attribute of each of the items in X.

A *grade assignment* G *from* $P = \langle X, \alpha \rangle$ *into* $\langle T, \succeq\ I \rangle$ assigns to each $x \in X$ a term $G(x) \in T$. A language measuring $\langle V, \geq \rangle$ is *suitable* for P if $\forall x \in X$, $\alpha(x) \in V$.

A *solution* to P is a pair $\langle G, L \rangle$, of which G is a grade assignment from P into L, and L is suitable for P. It determines a weak ordering (or ranking) of X in respect of α: x ranks higher than y if $G(x) \succeq G(y)$; x and y are tied in the implicit ranking if $G(x) = G(y)$.

> *Example:* Let H be the herd of oxen in the competition. The weight of any $x \in H$ in kilos, *weight*(x), is a positive real number. We face the problem $P = \langle H, \mathit{weight} \rangle$. Say you award each $x \in H$ either the grade h or an l. This fixes a grade assignment G_{you} from P into L_{you}. L_{you} interprets the grades in the positive reals and is suitable for P. So $\langle G_{\mathit{you}}, L_{\mathit{you}} \rangle$ is a solution to the ox-grading problem. Ox x ranks higher than ox y if $G_{\mathit{you}}(x) = h$ and $G_{\mathit{you}}(y) = l$.

Let $\langle G, \langle T, \succeq, I \rangle \rangle$ be a solution to $P = \langle X, \alpha \rangle$. It is an *accurate* solution to P *by its own standards* if for any $x \in X$, $\alpha(x) \in I(G(x))$. Intuitively,

x's grade is correct, given the truth about x and the meaning of the grade. Accurate solutions are desirable because their rankings tell the truth: if $G(x) \succ G(y)$ then $\alpha(x) > \alpha(y)$.[10] This follows from the orderliness of interpretations.

> *Example:* Your solution $\langle G_{you}, L_{you} \rangle$ to $\langle H, weight \rangle$ is accurate, by its own standards, if $G_{you}(x) = h$ for each $x \in H$ such that $weight(x) \geq 1000$, and $G_{you}(x) = l$ if $weight(x) < 1000$. Each ox gets whichever grade is correct, on your understanding of the grades.

4. Superlanguages and supersolutions

Grade languages make up languages of more-precise supergrades, and solutions to grading problems make up supersolutions in these superlanguages.

Suppose $L_i = \langle T_i, \succeq_i, I_i \rangle$ and $L_j = \langle T_j, \succeq_j, I_j \rangle$ are languages for $\langle V, \geq \rangle$. Let T_{ij} be $\{\langle e, f \rangle : e \in T_i, f \in T_j \text{ and } I_i(e) \cap I_j(f) \neq \emptyset\}$, and set $\langle e, f \rangle \succeq_{ij} \langle g, h \rangle$ if both $e \succeq_i g$ and $f \succeq_j h$. For each $\langle e, f \rangle \in T_{ij}$ let furthermore $I_{ij}\langle e, f \rangle$ be $I_i(e) \cap I_j(f)$. Define finally $L_i \circ L_j = \langle T_{ij}, \succeq_{ij}, I_{ij} \rangle$.

Lemma 1: $L_i \circ L_j$ is a grade language for $\langle V, \geq \rangle$.

There is a proof in the *Appendix*. $L_i \circ L_j$ is called the *superlanguage of L_i and L_j*.

> *Example:* Just by grading separately in L_{you} and L_{me}, the group $\langle you, me \rangle$ of us grades the oxen in $L_{you} \circ L_{me}$. Its vocabulary is $\{\langle h, h \rangle, \langle l, h \rangle, \langle l, l \rangle\}$, with $\langle h, h \rangle \succeq_{you,me} \langle l, h \rangle \succeq_{you,me} \langle l, l \rangle$.[11] The interpretation $I_{you,me}(\langle l, h \rangle)$ of $\langle l, h \rangle$, for instance, is $I_{you}(l) \cap I_{me}(h) = [500, 1000)$.[12]

[10] They do not in general tell the *whole* truth: sometimes $G(x) = G(y)$ though $\alpha(x) > \alpha(y)$.

[11] The sequence $\langle h, l \rangle$ is not a term in this collective vocabulary because $I_{you}(h) \cap I_{me}(l) = [1000, \infty) \cap [0, 500) = \emptyset$. This won't limit our ability to express ourselves coherently as a pair because it's logically impossible for an ox to be h by your standards and l by mine.

[12] Note that the order of the group is just a device for keeping track of who contributes which grade of the common vocabulary. The pair $\langle me, you \rangle$ supergrades in the different but equivalent language $L_{me} \circ L_{you}$, its term $\langle h, l \rangle$ denoting the same range $[500, 1000)$ of weights.

Superlanguages are grade languages in their own right. They too can be combined by supergrading. Define recursively $\circ(\langle L_1 \rangle) = L_1$ and:

$$\circ(\langle L_1, \ldots, L_{m+1} \rangle) = \circ(\langle L_1 \ldots L_m \rangle) \circ L_{m+1}.$$

Now, where $L_1, \ldots L_n$ are any grade languages for $\langle V, \geq \rangle$,

Theorem 2 (Existence of Superlanguages): $\circ(\langle L_1, \ldots L_n \rangle)$ is a grade language for $\langle V, \geq \rangle$.

The proof is a simple induction on n, with lemma 1 as induction step. Notice that the *composing languages* $L_1, \ldots L_n$ need not have anything in common, apart from measuring $\langle V, \geq \rangle$. They may have the same vocabularies or different ones. They may have few grade terms or many, independently of one another and under any interpretations at all.

> *Example:* You and I are joined in the ox-ranking competition by a newcomer who shares our binary vocabulary but gives it yet another interpretation: $I_{new}(h) = [750, \infty)$ and $I_{new}(l) = [0, 750)$. The group $\langle \mathit{you}, \mathit{me}, \mathit{new} \rangle$ uses superlanguage $\circ(\langle L_{you}, L_{me}, L_{new} \rangle)$ with vocabulary:
>
> $$\langle h, h, h \rangle \succeq_{you,me,new} \langle l, h, h \rangle \succeq_{you,me,new} \langle l, h, l \rangle \succeq_{you,me,new} \langle l, l, l \rangle.$$
>
> The interpretations of these terms are, from the top down, $[1000, \infty)$, $[750, 1000)$, $[500, 750)$ and $[0, 500)$.[13]

One language $\langle T_1, \succeq_1, I_1 \rangle$ is said to be *as precise as* another, $\langle T_2, \succeq_2, I_2 \rangle$, if for each $e \in T_2$ there is $T_e \subseteq T_1$ such that $I_2(e) = \bigcup \{ I_1(t) : t \in T_e \}$.

Fact 3: $\circ(\langle L_1, \ldots, L_n \rangle)$ is as precise as each of $L_1, \ldots, L_n$.d

There is a proof in the *Appendix.*

L_1 is (strictly) *more* precise than L_2 if L_1 is as precise as L_2 but L_2 is not as precise as L_1. If one language is as precise as another, and has a greater number of grades, it is more precise. This follows easily from the properties

[13] All but the outermost brackets of supergrade terms are to avoid clutter left out in this and coming examples. Really the top supergrade is $\langle \langle h, h \rangle, h \rangle$, and so on.

of interpretations. Superlanguages in particular are more precise than the composing languages when these have diverse standards.

Fact 4: Let $S_1, \ldots, S_n$ be sets of standards for several grade languages $L_1, \ldots, L_n$ that measure some common dimension. Then $\bigcup\{S_1, \ldots, S_n\}$ is a set of standards for $\circ(\langle L_1, \ldots, L_n\rangle)$.

There is a proof in the *Appendix.*

> *Example:* L_{you}, L_{me} and L_{new} have different standards for h. They are respectively 1000, 500 and 750. There is a common standard for l; it is 0. By fact 4, $\circ(\langle L_{you}, L_{me}, L_{new}\rangle)$ has supergrades with these four standards; they are respectively $\langle h,h,h\rangle$, $\langle l,h,l\rangle$, $\langle l,h,h\rangle$ and $\langle l,l,l\rangle$. By fact 3, $\circ(\langle L_{you}, L_{me}, L_{new}\rangle)$ is more precise than each of L_{you}, L_{me} and L_{new}, with just two grades.

Solutions to grading problems compose along with their languages. Consider two solutions $\langle G_i, L_i\rangle$ and $\langle G_j, L_j\rangle$ to $P = \langle X, \alpha\rangle$, each accurate by its own standards, of which the languages L_i and L_j measure the same dimension. Let $G_i \circ G_j$ map each $x \in X$ to the pair $\langle G_i(x), G_j(x)\rangle$. Then

Lemma 5: $\langle G_i \circ G_j, L_i \circ L_j\rangle$ is an accurate solution to P by its own standards.

There is a proof in the *Appendix.*

These *supersolutions* are solutions in their own right, and can be combined with other solutions. Let $\langle G_1, L_1\rangle, \ldots \langle G_n, L_n\rangle$ be solutions to some P, each accurate by its own standards, with $L_1, \ldots, L_n$ measuring some common dimension. Now define recursively a collective grade assignment, putting $\circ(\langle G_1\rangle) = G_1$ and:

$$\circ(\langle G_1, \ldots, G_{m+1}\rangle) = \circ(\langle G_1, \ldots, G_m\rangle) \circ G_{m+1}.$$

Then,

Theorem 6 (Existence of Accurate Supersolutions):

$$\langle \circ(\langle G_1, \ldots, G_n\rangle), \quad \circ(\langle L_1, \ldots, L_n\rangle)\rangle$$

is an accurate solution to P by its own standards.

The proof of theorem 6 is a simple induction on n, of which lemma 5 is the induction step.

> *Example:* Let $\langle G_{you}, L_{you}\rangle$, $\langle G_{me}, L_{me}\rangle$ and $\langle G_{new}, L_{new}\rangle$ be accurate solutions to $\langle H, weight\rangle$ by their own standards. By theorem 6 they make up a solution
>
> $$\langle \circ(\langle G_{you}, G_{me}, G_{new}\rangle), \quad \circ(\langle L_{you}, L_{me}, L_{new}\rangle)\rangle$$
>
> that is accurate by its own standards. Any given $x \in H$ receives the supergrade $\langle G_{you}(x), G_{me}(x), G_{new}(x)\rangle$.

5. From observations to grades

Hitting the bull's eye doesn't by itself make you a good shot. Anyone can be lucky. Similarly, there's more to ability in classification than somehow identifying the right classes for things: ability means *reliably* getting this right. This section characterizes reliability in grading as accuracy despite distortion of the signals on which grade decisions are based.

Consider some problem $P = \langle X, \alpha\rangle$. An individual i receives from each $x \in X$ a signal carrying information about $\alpha(x)$.[14] The collection φ of signals from all the X specifies, for each $x \in X$, some information $\varphi(x)$ about x. Here, for simplicity, each $\varphi(x)$ is a single value of the same sort that α takes. These are *generated* signals in the sense of Hong and Page [12], the "noisy glimpses or distortions of an outcome value" (p. 2177).[15]

An *individual scope* Φ for $\langle X, \alpha\rangle$ is the set of all such collections φ that some individual could receive from the X, depending on this individual's perspective or other circumstances under which the X are observed. An example gives intuitive content to the notion of an individual scope.

> *Example:* One fine day you observe the herd H and on one page of your notebook you write down a precise weight for each ox. This page of your notebook amounts to a collection of signals φ from H where, for any $x \in H$, $\varphi(x)$ is the weight you wrote down

[14] The term *signal* is used here to cover both conventional signals such as the level of applause at a public address and non-conventional *cues* such as the size of the visual image of an ox. What matters about them here is that they carry noisy information that is a basis for assigning grades.

[15] In a natural generalization, each $\varphi(x)$ is a collection of such values, such as the scattering of measurements read off an instrument or an imprecise interval.

> on this page for x. Under different circumstances – with different light, with the cattle arranged differently, observing from another distance or angle – you write down on another page somewhat different weights for each of the oxen in H. This second page is a different collection of signals, ψ. Your scope Φ_{you} for $\langle H, weight\rangle$ includes φ, ψ and all other collections of signals you might receive, on all of the different occasions on which you might possibly observe the herd of oxen. With a different page for each one, your scope is the whole notebook.

Consider more generally a group of people $1, \ldots n$ that is out to solve problem P. On some given occasion, each member i observes each $x \in X$ and thus receives some $\varphi_i \in \Phi_i$ from X; on this occasion the group $\langle 1, \ldots n\rangle$ receives signals $\vec{\varphi} = \langle \varphi_1, \ldots \varphi_n\rangle$. The group's *collective scope* for P is some $\Phi \subseteq \Phi_1 \times \ldots \times \Phi_n$. Just which such $\vec{\varphi}$ are in Φ depends on any connections between observations of different members of the group, on different possible occasions. The group is *independent* if $\Phi = \Phi_1 \times \ldots \times \Phi_n$.

The problem P at hand is solved by observing the items X and assigning to each one a grade from some suitable language $L = \langle T, \succeq, I\rangle$. A mapping $\mathcal{G} : \Phi \times X \to T$ is an *signaled-grade assignment from* Φ *and* P *into* L if for all $\vec{\varphi} \in \Phi$ and all $x, y \in X$, $\mathcal{G}(\vec{\varphi}, x) = \mathcal{G}(\vec{\varphi}, y)$ whenever, for every component φ_i of $\vec{\varphi}$, $\varphi_i(x) = \varphi_i(y)$. Intuitively, the grade for any $x \in X$ is fixed by everybody's signals just from x. Pairing such $\mathcal{G}$ with L we have a *signaled solution* $\langle \mathcal{G}, L\rangle$ *to* P *with* scope Φ.

Take any signaled solution $\langle \mathcal{G}, L\rangle$ to P. Holding fixed some $\vec{\varphi} \in \Phi$, its scope, we obtain a grade assignment $\mathcal{G}^{\vec{\varphi}}$ from P into L, defined by $\mathcal{G}^{\vec{\varphi}}(x) = \mathcal{G}(\vec{\varphi}, x)$. $\langle \mathcal{G}, L\rangle$ is a *reliably accurate solution to* P *with scope* Φ if for every $\vec{\varphi}$ in Φ, $\langle \mathcal{G}^{\vec{\varphi}}, L\rangle$ is an accurate solution to P, by its own standards. Where Φ is an individual scope we have, replacing $\vec{\varphi}$ by φ throughout, notions of an individual signaled solution and a reliably accurate individual solution.

Reliably accurate solutions are *reliable*: they invariably assign the same grades–correct ones, by their standards–despite variability in signals due to noise and distortion. This is easily seen. Let $\langle \mathcal{G}, L\rangle$ be any reliably accurate solution to P. Consider any $\vec{\varphi}, \vec{\psi} \in \Phi$, its scope. We have for any given $x \in X$, $\alpha(x) \in I(\mathcal{G}^{\vec{\varphi}}(x))$, and that is equivalent to $\mathcal{G}^{\vec{\varphi}}(x) = I^{-1}(\alpha(x))$. Similarly, $I^{-1}(\alpha(x)) = \mathcal{G}^{\vec{\psi}}(x)$. So $\mathcal{G}(\vec{\varphi}, x) = \mathcal{G}(\vec{\psi}, x)$.

Reliable accuracy can be achieved by using grades that are sufficiently coarse-grained. They have to mask variability in signals.

Consider first an individual scope Φ for the problem $P = \langle X, \alpha \rangle$ at hand. For any $x \in X$, let $\Phi(x)$ be $\{\varphi(x) : \varphi \in \Phi\}$, the possible signals from x. A language $\langle T, \succeq I \rangle$ *masks* Φ *in* P if for all $x \in X$ and all $e, f \in T$, if $\Phi(x) \cap I(e) \neq \emptyset$ and $\Phi(x) \cap I(f) \neq \emptyset$, then $e = f$. Intuitively, all possible signals from any given item are covered by a single grade. Say also that Φ is *truth compatible* for P if for each $x \in X$, $\alpha(x) \in \Phi(x)$. This just means that it is *possible* to observe the truth about the items.[16]

Now take some L that is *suitable* for Φ in P. That is, where $\langle V, \geq \rangle$ is the dimension that L measures, for every $\varphi \in \Phi$ and $x \in X$, let $\varphi(x) \in V$. Then no matter which signal is received from any given x in X, L has an applicable grade. Putting $\mathcal{G}_L(\varphi, x) = I^{-1}(\varphi(x))$, there is a signaled solution $\langle \mathcal{G}_L, L \rangle$ to P with scope Φ, and:

Fact 7: If individual scope Φ is truth compatible for P, then $\langle \mathcal{G}_L, L \rangle$ is a reliably accurate individual solution to P with scope Φ if and only if L masks Φ in P.

There is a proof in the *Appendix.*

Supergrading delivers reliably accurate groups. Given n reliably accurate individual solutions $\langle \mathcal{G}_i, L_i \rangle$ to P with scopes Φ_i, all L_i measuring some common dimension, we can define by recursion a signaled grade assignment $\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_n \rangle)$ from $\Phi_1 \times \ldots \times \Phi_n$ and P into $\circ(\langle L_1, \ldots, L_n \rangle)$. For any $\varphi_i \in \Phi_i$ and $x \in X$, put $\circ(\langle \mathcal{G}_1 \rangle)(\langle \varphi_1 \rangle, x) = \mathcal{G}_1(\varphi_1, x)$ and:

$$\begin{gathered}\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1} \rangle)(\langle \varphi_1, \ldots, \varphi_{m+1} \rangle, x) = \\ \langle \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m \rangle)(\langle \varphi_1, \ldots, \varphi_m \rangle, x), \quad \mathcal{G}_{m+1}(\varphi_{m+1}, x) \rangle.\end{gathered}$$

Then,

Theorem 8 (Existence of Reliably Accurate Supersolutions):

$$\langle \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_n \rangle), \quad \circ(\langle L_1, \ldots, L_n \rangle) \rangle$$

is a reliably accurate solution to P with scope $\Phi_1 \times \ldots \times \Phi_n$.

There is a proof in the *Appendix.*

[16] It could for example happen that you put an ox on a livestock scale and read off its true weight.

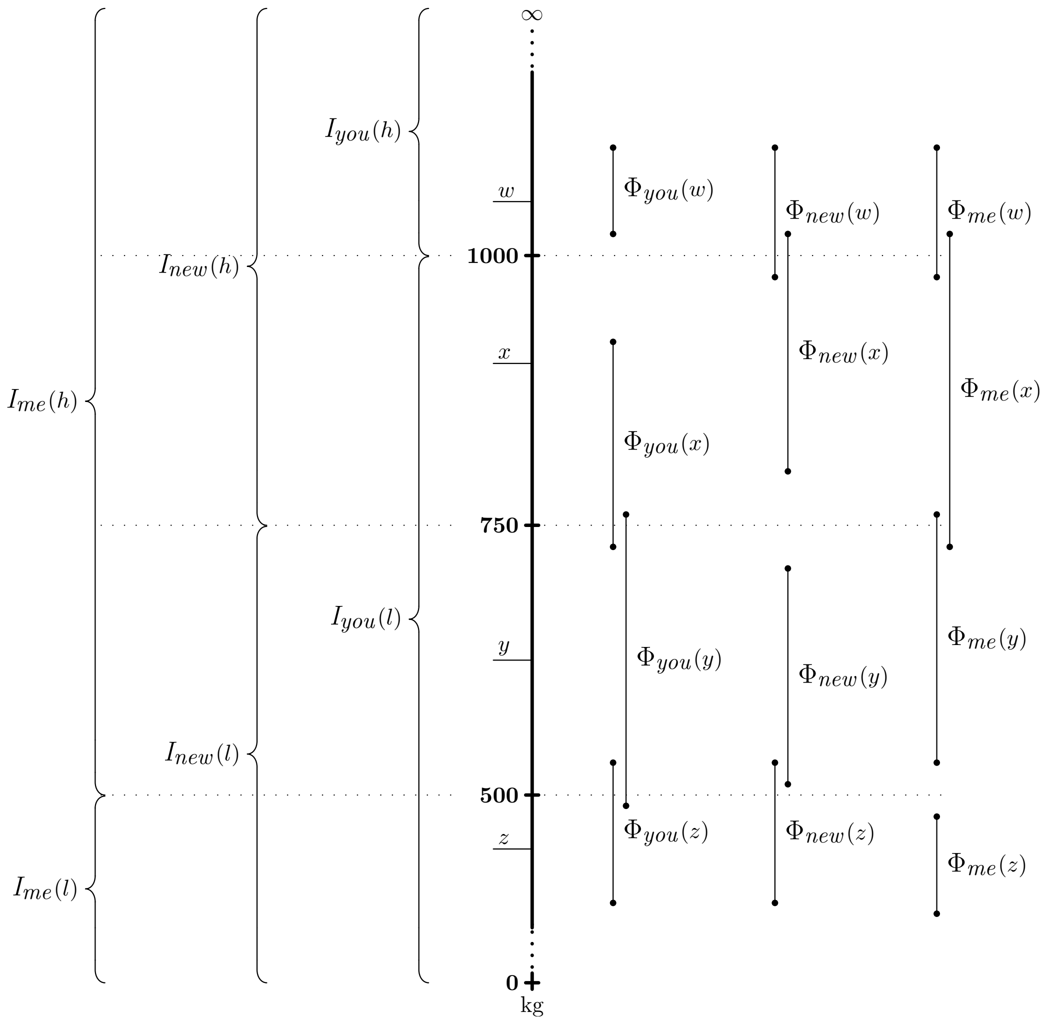

Figure 1: Individual scope Φ_{you} is truth compatible and masked by L_{you}. Similarly for Φ_{me} and L_{me}, and for Φ_{new} and L_{new}. The individual ability in $\langle H, weight\rangle$ of three graders with these scopes is just $\frac{2}{4}$. The collective ability is a perfect $\frac{4}{4}$.

> Example: Assume now that H includes just four oxen w, x, y and z, with $weight(w) = 1125$, $weight(x) = 875$, $weight(y) = 625$ and $weight(z) = 375$. Let Φ_{you}, Φ_{me} and Φ_{new} be as in Fig. 1. By inspection, these scopes are truth compatible; also, L_{you}, L_{me} and L_{new} mask them. There are by fact 7 reliably accurate individual solutions $\langle \mathcal{G}_{you}, L_{you} \rangle$, $\langle \mathcal{G}_{me}, L_{me} \rangle$ and $\langle \mathcal{G}_{new}, L_{new} \rangle$ with these scopes, and by theorem 8 there is a reliably accurate collective solution $\langle \circ(\langle \mathcal{G}_{you}, \mathcal{G}_{me}, \mathcal{G}_{new}, \rangle), \quad \circ(\langle L_{you}, L_{me}, L_{new} \rangle) \rangle$ with scope $\Phi_{you} \times \Phi_{me} \times \Phi_{new}$.

6. Ability = accuracy+reliability+*discrimination*

Even *reliably* hitting the bulls eye doesn't make you a good shot. It might be very big. Similarly, ability in solving a grading problem means more than reliably assigning correct grades, which might depending on the language be very imprecise. The grades have to tell different items apart.

Consider an accurate solution $\langle G, L \rangle$ to $P = \langle X, \alpha \rangle$. Its *discrimination* in P is $|\{e : \exists x \in X, G(x) = e\}|$. This is the number of equivalence classes into which G divides X. The discrimination of a reliably accurate solution $\langle \mathcal{G}, L \rangle$ with scope Φ is that of $\langle \mathcal{G}^{\vec{\varphi}}, L \rangle$ for any (and all) $\vec{\varphi}$ in Φ.

> *Example:* Let $\langle \mathcal{G}_{you}, L_{you} \rangle$ be a reliably accurate solution to $\langle H, weight \rangle$ with scope Φ_{you}, from Fig. 1. For each $\varphi \in \Phi_{you}$, $\mathcal{G}^{\varphi}_{you}(w) = h$ and $\mathcal{G}^{\varphi}_{you}(x) = \mathcal{G}^{\varphi}_{you}(y) = \mathcal{G}^{\varphi}_{you}(z) = l$. The discrimination of $\langle \mathcal{G}_{you}, L_{you} \rangle$ in $\langle H, weight \rangle$ is 2. Reliably accurate $\langle \mathcal{G}_{me}, L_{me} \rangle$ and $\langle \mathcal{G}_{new}, L_{new} \rangle$ have the same discrimination.

When one and the same signal could be received from either of two items it is not possible to tell them apart by reliably and accurately assigning them different grades. Let us say that x and y are *compatible in* individual scope Φ if for any $\varphi, \psi \in \Phi$ there is some $\chi \in \Phi$ such that $\chi(x) = \varphi(x)$ and $\chi(y) = \psi(y)$. Intuitively, if some given signals can at all be received from the two items, perhaps on different occasions, then these two signals can also be received from them together, on the same occasion. We have:

Theorem 9 (Limits to Discrimination): Let $\langle \mathcal{G}, L \rangle$ be a reliably accurate solution to $\langle X, \alpha \rangle$ with scope $\Phi_1 \times \ldots \times \Phi_n$. Let $x \in X$ and $y \in X$ be such that $\Phi_i(x) \cap \Phi_i(y) \neq \emptyset$, for each $1 \leq i \leq n$. Let x and y be compatible in each Φ_i. Then for each $\vec{\varphi} \in \Phi$, $\mathcal{G}(\vec{\varphi}, x) = \mathcal{G}(\vec{\varphi}, y)$.

There is a proof in the *Appendix*. It makes clear that with $n = 1$ the corresponding result for reliably accurate *individual* solutions is a special case.

Ability is now a matter of the discrimination that *can* be achieved, despite the noise and distortion in signals. Let the *maximal discrimination* of Φ in P be the discrimination of any solution which, among all reliably accurate solutions to P with scope Φ, has the greatest discrimination.[17] The *ability* of a person or group in $\langle X, \alpha \rangle$ is now the maximal discrimination of their Φ divided by the number of equivalence classes into which α partitions X. It is the best that can be done expressed as a proportion of what may at all be hoped for.

> *Example:* Let each of $\Phi_{you}(x)$, $\Phi_{you}(y)$ and $\Phi_{you}(z)$ overlap the next, as in Fig. 1. Let x and y be compatible in Φ_{you}, as are y and z. Let $\langle \mathcal{G}, L \rangle$ be *any* reliably accurate solution to $P = \langle H, weight \rangle$ with scope Φ_{you}, using any language L at all. By theorem 9, for any $\varphi \in \Phi_{you}$, $\mathcal{G}(\varphi, x) = \mathcal{G}(\varphi, y) = \mathcal{G}(\varphi, z)$. Therefore no reliably accurate solution with this scope has greater discrimination than 2. There is a solution using L_{you} with discrimination 2, so this is the maximum discrimination of Φ_{you} in P. The oxen divide into 4 classes by weight, so your ability in P is $\frac{2}{4}$. My ability and the newcomer's are under similar assumptions also $\frac{2}{4}$.

7. Diversity in grading standards beats individual ability

A group with diverse standards can have greater collective ability than a less diverse group whose individual members have greater ability. Two examples illustrate.

> *Example:* Consider the group of Fig.1, with any collective scope $\Phi \subseteq \Phi_{you} \times \Phi_{me} \times \Phi_{new}$. The previous example shows that the individual abilities in $P = \langle H, weight \rangle$ are $\frac{2}{4}$. The group by fact 7 and theorem 8 has a reliably accurate solution $\langle \mathcal{G}, L \rangle$ with scope Φ, where $L = \circ(\langle L_{you}, L_{me}, L_{new} \rangle)$. Taking any $\vec{\varphi} \in \Phi$, $\langle \mathcal{G}^{\vec{\varphi}}, L \rangle$ is an accurate solution to P, by its own standards, and assigns to each ox in H the correct supergrade. Thus,

[17] Assuming X is finite there has to be a maximum; let it be so.

$$\mathcal{G}^{\vec{\varphi}}(w) = \langle h, h, h \rangle.$$

At 1125 kilos, w meets the standard in L for this grade, which is 1000 kilos, and there is no higher grade. Also,

$$\mathcal{G}^{\vec{\varphi}}(x) = \langle l, h, h \rangle,$$

since x at 875 kilos meets the standard of 750 for this grade, but falls short of 1000, for the next. Similarly:

$$\mathcal{G}^{\vec{\varphi}}(y) = \langle l, h, l \rangle \text{ (since } 500 \leq 625 < 750\text{) and}$$
$$\mathcal{G}^{\vec{\varphi}}(z) = \langle l, l, l \rangle \text{ (since } 0 \leq 375 < 500\text{).}$$

All 4 supergrades are assigned. The discrimination of $\langle \mathcal{G}, L \rangle$ in P is 4. That's as many as there are (equivalence classes by weight of) oxen in P—as good as can be. The collective ability in P is a perfect $\frac{4}{4}$. This group has just the heaviest ox w in its top weight category and wins the competition.

Example: Three graders have the same scope Φ, as seen in Fig. 2. By inspection it is truth compatible for $P = \langle H, weight \rangle$, and $\Phi(w)$ overlaps $\Phi(x)$. Assume furthermore that w and x are compatible in Φ,[18] and that the three form an independent group, with collective scope Φ^3.

By theorem 9, any reliably accurate individual or collective solution with scope Φ or Φ^3 must assign w and x the same grade. The maximal discrimination in P is therefore at best 3. Choosing a common ternary language that masks Φ (see Fig. 2), fact 7 and theorem 8 provide reliably accurate individual and collective solutions with scopes Φ and Φ^3 whose discrimination is 3. The individual and collective abilities in P are $\frac{3}{4}$.

This group, despite the greater individual ability of its members, has less collective ability in P than the more-diverse first group. Besides w it has also the second-heaviest ox x in the top category by weight. Choosing at random between them, this group only has even odds of winning.

[18] The example easily generalizes so that there are distinct but somewhat similar scopes satisfying these conditions.

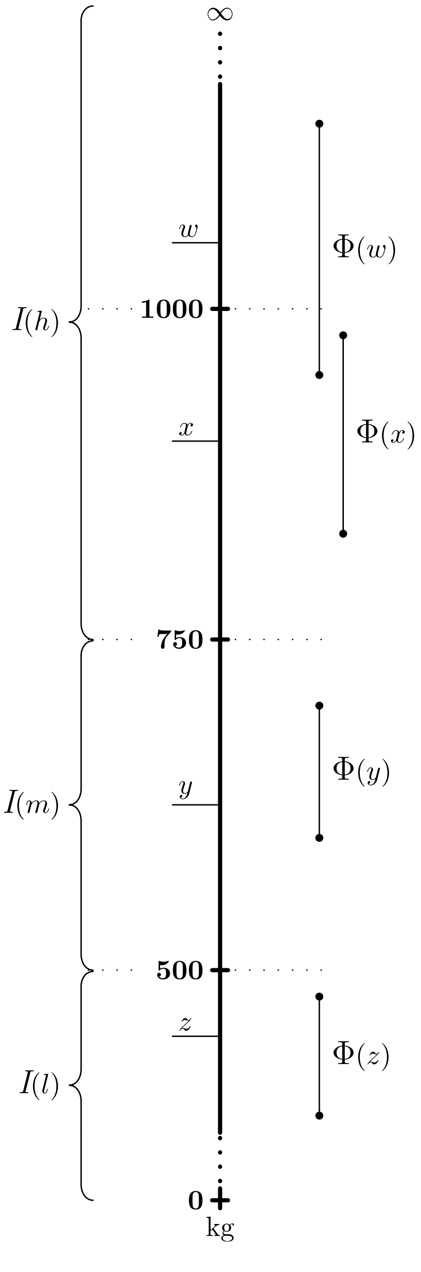

Figure 2: Individual scope Φ is truth compatible and masked by a language with 3 grades h, m and l. Graders with this scope have individual ability $\frac{3}{4}$ in $\langle H, weight\rangle$, higher than with the first group. But the collective ability, also $\frac{3}{4}$, is with the lack of diversity lower.

8. Concluding remarks

These findings suggest that in collective judgment it might often be a good idea to use inputs that are not very precise. This can make it easier to gather accurate information from diverse sources. Mismatched meanings can later be exploited to achieve through aggregation the ultimately desired precision and informativeness.

That inconsistent grading standards are an asset might surprise organizers and members of committees and expert panels who feel that everybody should be “on the same page”. In ordinary cooperative conversation, after all, it is proper to avoid ambiguity and equivocation [11]. One upshot of the results presented here is when the goal of the conversation is to exchange accurate information for purposes of collective judgment it might sometimes be better to suspend this norm.

Condorcet’s Jury Theorem underlines the advantage in drawing from independent sources of information [7]. Hong and Page have shown how diverse heuristics and representations can help with search problems [13]. These findings add understandings of language to the list of differences that can give diverse groups an edge in gathering knowledge.

Much of interest has been left for future work. A referee conjectures that if players choose their thresholds simultaneously then, provided they care about collective ability, optimal thresholds like those of Fig. 1 are the Nash equilibria of the game defined over selecting thresholds. If on the other hand players care about individual ability, then suboptimal thresholds like those of Fig. 2 will be equilibria, and incentives for diversity will be needed in order to promote collective ability. It would be good to have a full characterization of the optimal grading thresholds for any given classification problem of the kind considered here, in order to understand the conditions under which high collective ability may be expected to develop.

There are ramifications in machine learning. In learning by Random Forest, many tree-structured classifiers are grown from random selections of the same training data; voting among these tree classifiers as to which is the right class is known to bring about a significant improvement in accuracy of the forests in comparison with the individual trees [4]. Where the classes of interest are intervals of a common underlying dimension, as is the case with scoring and grading, supergrading could be used to form superforests, or ensembles of random forests. In this way, classifiers could be obtained that like the constitutive random forests are highly accurate, but which classify items more precisely than they do.

[1] Balinski, M. and R. Laraki (2007). A theory of measuring, electing, and ranking. *Proc. Natl. Acad. Sci. USA 104*(21), 8720–8725.

[2] Balinski, M. and R. Laraki (2011). *Majority Judgement.* Cambridge MA: MIT Press.

[3] Balshem, H., M. Helfand, H. J. Schunemann, 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.

[4] Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

[5] Budescu, D. V., S. Broomell, and H.-H. Por (2009). Improving communication of uncertainty in the reports of the Intergovernmental Panel on Climate Change. *Psychological Science 20*(3), 299–308.

[6] Budescu, D. V., H.-H. Por, S. B. Broomell, and M. Smithson (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change 4*, 508–512.

[7] 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.

[8] Érdi, P. (2019). *Ranking: The Unwritten Rules of the Social Game We All Play.* Oxford University Press.

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

[10] Gottlieb, K. and F. Hussain (2015). Voting for image scoring and assessment (VISA) - theory and application of a 2+1 reader algorithm to improve accuracy of imaging endpoints in clinical trials. *BMC Medical Imaging 15*(1).

[11] Grice, P. (1989). *Studies in the Way of Words.* Cambridge MA: Harvard University Press.

[12] Hong, L. and S. Page (2009). Interpreted and generated signals. *Journal of Economic Theory 144*(5), 2174 – 2196.

[13] Hong, L. and S. E. Page (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci. USA 101*(46), 16385–16389.

[14] King, G., C. Murray, J. Salomon, and A. Tandon (2009). Enhancing the validity and cross-cultural comparability of measurement in survey research. In S. Pickel, G. Pickel, H.-J. Lauth, and D. Jahn (Eds.), *Methoden der vergleichenden Politik- und Sozialwissenschaft*, pp. 317–346. VS Verlag für Sozialwissenschaften.

[15] 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.

[16] Morgan, M. G. (1998). Uncertainty analysis in risk assessment. *Human and Ecological Risk Assessment 4*(1), 25–39.

[17] 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.

[18] 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.

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

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

[21] Wallsten, T. S., D. V. Budescu, A. Rapoport, R. Zwick, and B. Forsyth (1986). Measuring the vague meanings of probability terms. *J Exp Psychol Gen 155*(4), 348–365.

[22] 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.

Appendix

This section has proofs for technical claims in the main text, after repeating definitions of relevant notions introduced there.

Lemma 1

Let L_i and L_j be grade languages for $\langle V, \geq \rangle$. Then $L_i \circ L_j$ is a grade language for $\langle V, \geq \rangle$.

Definitions. $\langle T, \succeq, I \rangle$ is a *grade language* for $\langle V, \geq \rangle$ if $\langle T, \succeq \rangle$ is a *grade vocabulary* (that is, $T \neq \emptyset$ is finite and $\succeq$ is a linear ordering of T); and $I : T \to \wp(V)$ is an *interpretation* of $\langle T, \succeq \rangle$ in $\langle V, \geq \rangle$, that is:

- $\forall e \in T$, $I(e)$ is *convex*: $\forall u, v, w \in V$, if $u, w \in I(e)$ and $u \geq v \geq w$, then $v \in I(e)$;
- *I partitions V*: $\forall e \in T, I(e) \neq \emptyset$; $\forall e, f \in T$, if $I(e) \cap I(f) \neq \emptyset$ then $e = f$; and $\bigcup \{I(e) : e \in T\} = V$; and
- I is *orderly*: $\forall e, f \in T$, if $e \succ f$ then $I(e) > I(f)$ (that is: $\forall u \in I(e)$ and $\forall v \in I(f)$, $u > v$).

Where $L_i = \langle T_i, \succeq_i, I_i \rangle$ and $L_j = \langle T_j, \succeq_j, I_j \rangle$ are grade languages for a common $\langle V, \geq \rangle$, define:

- $T_{ij} = \{\langle e, f \rangle : e \in T_i, f \in T_j \text{ and } I_i(e) \cap I_j(f) \neq \emptyset\}$,
- $\langle e, f \rangle \succeq_{ij} \langle g, h \rangle$ if both $e \succeq_i g$ and $f \succeq_j h$,
- $I_{ij}\langle e, f \rangle = I_i(e) \cap I_j(f)$, for any $\langle e, f \rangle \in T_{ij}$, and
- $L_i \circ L_j = \langle T_{ij}, \succeq_{ij}, I_{ij} \rangle$.

Proof of lemma 1. To be shown are that (1) $\langle T_{ij}, \succeq_{ij} \rangle$ is a grade vocabulary and (2) I_{ij} is an interpretation of $\langle T_{ij}, \succeq_{ij} \rangle$ in $\langle V, \geq \rangle$.

(1) $T_{ij} \neq \emptyset$ since $\langle I_i^{-1}(v), I_j^{-1}(v) \rangle \in T_{ij}$, for any $v \in V$. ($I_i^{-1}(v)$ is the unique $e \in T_i$ such that $v \in I_i(e)$, and similarly for $I_j^{-1}(v)$.) T_{ij} is finite since T_i, T_j are. To see that $\succeq_{ij}$ is a linear ordering of T_{ij}, note that antisymmetry and transitivity of $\succeq_{ij}$ follow immediately from the corresponding properties of $\succeq_i$ and $\succeq_j$. To see that $\succeq_{ij}$ is a total ordering of T_{ij}, suppose $\langle e, f \rangle \not\succeq_{ij} \langle g, h \rangle$, i.e. either $e \not\succeq_i g$ or $f \not\succeq_j h$. Take the first case (the two are analogous).

Then, since $\succeq_i$ is total, $g \succ_i e$. Since $\langle g, h \rangle, \langle e, f \rangle \in T_{ij}$ it is possible to choose $u \in I_i(g) \cap I_j(h)$ and $v \in I_i(e) \cap I_j(f)$. Since I_i is orderly, $u > v$. Since I_j is orderly, $f \not\succ_j h$, and so, because $\succeq_j$ is total, $h \succeq_j f$. But then, as required, $\langle g, h \rangle \succeq_{ij} \langle e, f \rangle$.

(2) To be shown are that (i) $\forall \langle e, f \rangle \in T_{ij}, I_{ij}(\langle e, f \rangle)$ is convex; that (ii) I_{ij} partitions V; and that (iii) I_{ij} is orderly.

(i) Suppose $u, v, w \in V$, $u, w \in I_{ij}(\langle e, f \rangle)$ and $u \geq v \geq w$. By definition of I_{ij}, $u, w \in I_i(e) \cap I_j(f)$. Since both $I_i(e)$ and $I_j(f)$ are convex, $v \in I_i(e) \cap I_j(f) = I_{ij}(\langle e, f \rangle)$.

(ii) Consider any $\langle e, f \rangle, \langle g, h \rangle \in T_{ij}$. First, $I_{ij}(\langle e, f \rangle) \neq \emptyset$ by choice of the vocabulary T_{ij}; second, if $I_{ij}(\langle e, f \rangle) \cap I_{ij}(\langle g, h \rangle) \neq \emptyset$ then, expanding using the definition of I_{ij}, we have $I_i(e) \cap I_i(g) \neq \emptyset$, and also $I_j(f) \cap I_j(h) \neq \emptyset$. Because L_i and L_j are grade languages for $\langle V, \geq \rangle$, I_i partitions V and so does I_j. So $e = g$, $f = h$ and, as required, $\langle e, f \rangle = \langle g, h \rangle$; finally, $\bigcup \{ I_{ij}(\langle e, f \rangle) : \langle e, f \rangle \in T_{ij} \} = V$ because for any given $v \in V$, $\langle I_i^{-1}(v), I_j^{-1}(v) \rangle \in T_{ij}$ and $v \in I_i(I_i^{-1}(v)) \cap I_j(I_j^{-1}(v)) = I_{ij}(\langle I_i^{-1}(v), I_j^{-1}(v) \rangle)$.

(iii) Consider any $\langle e, f \rangle$, $\langle g, h \rangle \in T_{ij}$. Suppose $\langle e, f \rangle \succ_{ij} \langle g, h \rangle$. Consider any $u \in I_{ij}(\langle e, f \rangle)$ and $v \in I_{ij}(\langle g, h \rangle)$. Since $\langle g, h \rangle \not\succeq_{ij} \langle e, f \rangle$, either $g \not\succeq_i e$ or $h \not\succeq_j f$. $\succeq_i$ and $\succeq_j$ are total relations, so either $e \succ_i g$ or $f \succ_j h$. Suppose $e \succ_i g$. By definition of I_{ij}, $u \in I_i(e)$ and $v \in I_i(g)$. Since I_i is orderly, $u > v$. This follows similarly in case $f \succ_j h$. So $I_{ij}(\langle e, f \rangle) > I_{ij}(\langle g, h \rangle)$. □

Fact 3

$\circ(\vec{L})$ is as precise as each of its composing languages.

Definitions. Let languages $\vec{L} = \langle L_1, \ldots L_n \rangle$ measure some common dimension. Their superlanguage $\circ(\vec{L})$ is defined recursively: $\circ(\langle L_1 \rangle) = L_1$; $\circ(\langle L_1, \ldots, L_{m+1} \rangle) = \circ(\langle L_1 \ldots L_m \rangle) \circ L_{m+1}$. The *composing languages* are $L_1, \ldots L_n$. $\langle T_1, \succeq_1, I_1 \rangle$ is *as precise as* $\langle T_2, \succeq_2, I_2 \rangle$ if for each $e \in T_2$ there is $T_e \subseteq T_1$ such that $I_2(e) = \bigcup \{ I_1(t) : t \in T_e \}$.

Proof of fact 3. An induction on the length of $\vec{L}$, of which the induction step uses: *Lemma.* Let L_i, L_j and $L_i \circ L_j$ be as in the statement of lemma 1. Then $L_i \circ L_j$ is as precise as L_i and $L_i \circ L_j$ is as precise as L_j. *Proof of the lemma.* To be shown is that $L_i \circ L_j$ is as precise as L_i. (The similar demonstration that it is as precise as L_j is omitted.) Required is that for any $e \in T_i$ there is $T_e \subseteq T_{ij}$ such that $I_i(e) = \bigcup \{ I_{ij}(t) : t \in T_e \}$. For any given $e \in T_i$ set $T_e = \{ \langle e, f \rangle : \langle e, f \rangle \in T_{ij} \}$. That $I_i(e) \supseteq \bigcup \{ I_{ij}(t) : t \in T_e \}$ is easily seen since for any $\langle e, f \rangle \in T_{ij}$ we have $I_i(e) \supseteq I_i(e) \cap I_j(f) = I_{ij}(\langle e, f \rangle)$. To see that furthermore $I_i(e) \subseteq \bigcup\{I_{ij}(t) : t \in T_e\}$, consider any $v \in I_i(e)$. Let f be $I_j^{-1}(v)$. Then $v \in I_i(e) \cap I_j(f)$. So $\langle e, f\rangle \in T_{ij}$, and $\langle e, f\rangle \in T_e$. Now $v \in I_i(e) \cap I_j(f) = I_{ij}(\langle e, f\rangle) \subseteq \bigcup\{I_{ij}(t) : t \in T_e\}$. This completes the proof of the lemma.

Note before the proof of fact 3 itself that comparative precision is both reflexive (L is as precise as L) and transitive (if L_1 is as precise as L_2, and L_2 as precise as L_3, then L_1 is as precise as L_3. Now we have:

Base step: $\vec{L} = \langle L_1\rangle$. Then $\circ(\vec{L}) = L_1$. It is as precise as the only composing language, L_1, because comparative precision is reflexive.

Inductive step: $\vec{L} = \langle L_1, \ldots, L_{m+1}\rangle$. By the lemma, $\circ(\vec{L}) = \circ(\langle L_1, \ldots, L_m\rangle) \circ L_{m+1}$ is as precise as $\circ(\langle L_1, \ldots, L_m\rangle)$. By induction hypothesis, $\circ(\langle L_1, \ldots, L_m\rangle)$ is as precise as each of $L_1, \ldots, L_m$. By transitivity of comparative precision therefore $\circ(\vec{L})$ is as precise as each of $L_1, \ldots, L_m$. By the lemma, furthermore, $\circ(\vec{L})$ is as precise as the single remaining composing language, L_{m+1}. □

Fact 4

Let $S_1, \ldots, S_n$ be sets of standards for languages $L_1, \ldots, L_n$. Then $\bigcup\{S_1, \ldots, S_n\}$ is a set of standards for the superlanguage $\circ(\langle L_1, \ldots, L_n\rangle)$.

Definitions. Let $L = \langle T, \succeq, I\rangle$ be a language for $\langle V, \geq\rangle$. s is the *standard* for e *in* L if:

- $e \in T$,
- $s \in I(e)$,
- $I^{-1}(v) \succeq e$ for any $v \in V$ such that $v \geq s$, and
- $I^{-1}(v) \prec e$ for any $v \in V$ such that $v < s$.

S is a *set of standards* for L if for each $s \in S$ there is some $e \in T$ such that s is the standard for e in L.

Proof of fact 4. An induction on n, of which the induction step uses the following *Lemma.* Let L_i and L_j be languages for $\langle V, \geq\rangle$. Let S_i be a set of standards for L_i and let S_j be a set of standards for L_j. Then $S_i \cup S_j$ is a set of standards for $L_i \circ L_j$. *Proof of the lemma.* Consider any $s \in S_i \cup S_j$. There are two cases to consider: $s \in S_i$ and $s \in S_j$. We consider just the first case (the argument for the second case is a mirror image of that for the first). With $s \in S_i$, let $e \in T_i$ be such that s is the standard for e in L_i.

We will see that s is the standard for $\langle e, I_j^{-1}(s)\rangle$ in L_{ij}. Required are (1) $\langle e, I_j^{-1}(s)\rangle \in T_{ij}$, (2) $s \in I_{ij}(\langle e, I_j^{-1}(s)\rangle)$, (3) $I_{ij}^{-1}(v) \succeq_{ij} \langle e, I_j^{-1}(s)\rangle$ for any $v \in V$ such that $v > s$, and (4) $I_{ij}^{-1}(v) \prec_{ij} \langle e, I_j^{-1}(s)\rangle$ for any $v \in V$ such that $v < s$.

(1) Because s is the standard for e in L_i, $s \in I_i(e)$. Furthermore, $s \in I_j(I_j^{-1}(s))$, so $I_i(e) \cap I_j(I_j^{-1}(s)) \neq \emptyset$ and $\langle e, I_j^{-1}(s)\rangle \in T_{ij}$.

(2) $s \in I_i(e) \cap I_j(I_j^{-1}(s))$ as in (1). Now, by definition of I_{ij}, $I_i(e) \cap I_j(I_j^{-1}(s)) = I_{ij}(\langle e, I_j^{-1}(s)\rangle)$.

(3) Notice first that for any $v \in V$ we have $I_{ij}^{-1}(v) = \langle I_i^{-1}(v), I_j^{-1}(v)\rangle$. This is equivalent to $v \in I_{ij}(\langle I_i^{-1}(v), I_j^{-1}(v)\rangle)$, which follows by unpacking the relevant definitions. Now, consider any $v \in V$ such that $v > s$. Since both I_i and I_j are orderly, $I_i^{-1}(v) \succeq_i I_i^{-1}(s)$ and also $I_j^{-1}(v) \succeq_j I_j^{-1}(s)$. So as required $I_{ij}^{-1}(v) = \langle I_i^{-1}(v), I_j^{-1}(v)\rangle \succeq_{ij} (I_i^{-1}(s), I_j^{-1}(s)) = \langle e, I_j^{-1}(s)\rangle$. The last equality is due to the fact that since s is the standard for e in L_i we have $s \in I_i(e)$ or, equivalently, $I_i^{-1}(s) = e$.

(4) Consider any $v \in V$ such that $v < s$. Because s is the standard for e in L_i, $I_i^{-1}(v) \prec_i e = I_i^{-1}(s)$, so $\langle I_i^{-1}(v), I_j^{-1}(v)\rangle \not\succeq_{ij} (I_i^{-1}(s), I_j^{-1}(s))$. Since $\succeq_{ij}$ is a total ordering of T_{ij} we have as required $I_{ij}^{-1}(v) = \langle I_i^{-1}(v), I_j^{-1}(v)\rangle \prec_{ij} (I_i^{-1}(s), I_j^{-1}(s)) = \langle e, I_j^{-1}(s)\rangle$. This completes the proof of the lemma, and of fact 4. □

Lemma 5

Let $\langle G_i, L_i\rangle$ and $\langle G_j, L_j\rangle$ be solutions to problem P, each accurate by its own standards, where L_i and L_j measure a common dimension. Then

$$\langle G_i \circ G_j, \quad L_i \circ L_j\rangle$$

is an accurate solution to P, by its own standards.

Definitions. Let $L = \langle T, \succeq, I\rangle$ be a grade language for $\langle V, \geq\rangle$ and let $P = \langle X, \alpha\rangle$ be a grading problem. Then

- L is *suitable* for P if $\forall x \in X$, $\alpha(x) \in V$,
- $\langle G, L\rangle$ is a *solution* to P if G is an assignment from P into L (that is, $\forall x \in X, G(x) \in T$) and L is suitable for P,
- $\langle G, L\rangle$ is an *accurate solution to P, by its own standards* if $\langle G, L\rangle$ is a solution to P and $\forall x \in X$, $\alpha(x) \in I(G(x))$.

Furthermore,

- T_{ij}, I_{ij} and $L_i \circ L_j$ are as in the superlanguage lemma,

and for mappings $G_i : X \to T_i$ and $G_j : X \to T_j$, we define $G_i \circ G_j : X \to T_i \times T_j$ by putting, for each $x \in X$,

- $G_i \circ G_j(x) = \langle G_i(x), G_j(x) \rangle$.

Proof of lemma 5. To be shown are that (1) $\langle G_i \circ G_j, \quad L_i \circ L_j \rangle$ is a solution to P, and (2) for each $x \in X$, $\alpha(x) \in I_{ij}(G_i \circ G_j(x))$.

(1) Required are that (i) $G_i \circ G_j$ maps each $x \in X$ to some term in T_{ij}, the grade terms of $L_i \circ L_j$, and (ii) $L_i \circ L_j$ is suitable for P.

(i) Consider any $x \in X$. Since $\langle G_i, L_i \rangle$ and $\langle G_j, L_j \rangle$ are accurate by their own standards, both $\alpha(x) \in I_i(G_i(x))$ and $\alpha(x) \in I_j(G_j(x))$. So $\alpha(x) \in I_i(G_i(x)) \cap I_j(G_j(x)) \neq \emptyset$. Therefore, by definition of $G_i \circ G_j$ and of T_{ij}, $G_i \circ G_j(x) = \langle G_i(x), G_j(x) \rangle \in T_{ij}$.

(ii) By lemma 1, $L_i \circ L_j$ measures the same dimension as L_i (and L_j). Suitability of $L_i \circ L_j$ for P follows immediately from that of L_i (or that of L_j).

(2) Consider any $x \in X$. As in (1) (i), $\alpha(x) \in I_i(G_i(x))$ and $\alpha(x) \in I_j(G_j(x))$. So, by definition of I_{ij} and $G_i \circ G_j$, $\alpha(x) \in I_i(G_i(x)) \cap I_j(G_j(x)) = I_{ij}(\langle G_i(x), G_j(x) \rangle) = I_{ij}(G_i \circ G_j(x))$. □

Fact 7

Let individual scope Φ be truth compatible for P, and let L be suitable for Φ in P. Then $\langle \mathcal{G}_L, L \rangle$ is a reliably accurate solution to P with scope Φ if and only if L masks Φ in P.

Definitions. Let $L = \langle T, \succeq, I \rangle$ measure $\langle V, \geq \rangle$, and $P = \langle X, \alpha \rangle$. Let Φ be an individual scope, and let $\langle \mathcal{G}, L \rangle$ be an individual signaled solution to P with scope Φ. Define

- L is *suitable for* P if $\forall x \in X, \alpha(x) \in V$,
- L is *suitable for* Φ *in* P if $\forall \varphi \in \Phi, \forall x \in X$, $\varphi(x) \in V$,
- $\mathcal{G}_L : \Phi \times X \to T$ is the signaled grade assignment from Φ and P into L such that $\mathcal{G}_L(\varphi, x) = I^{-1}(\varphi(x))$ (defined only in case L is suitable for Φ in P),

- $\Phi(x) = \{\varphi(x) : \varphi \in \Phi\}$, for any given $x \in X$,
- L *masks* Φ *in* P if for all $x \in X$ and all $e, f \in T$, if $\Phi(x) \cap I(e) \neq \emptyset$ and $\Phi(x) \cap I(f) \neq \emptyset$, then $e = f$,
- Φ is *truth compatible* for P if for each $x \in X$, $\alpha(x) \in \Phi(x)$,
- A signaled solution $\langle \mathcal{G}, L\rangle$ to P with scope Φ is a *reliably accurate solution to* P with scope Φ if for every $\varphi \in \Phi$, $\langle \mathcal{G}^{\varphi}, L\rangle$ is an accurate solution to P, by its own standards.

Proof of fact 7. For the *if* half, consider under the assumptions of the theorem the signaled solution $\langle \mathcal{G}_L, L\rangle$ to P with scope Φ. Suppose $L = \langle T, \succeq, I\rangle$ masks Φ in $P = \langle X, \alpha\rangle$. Take any $\varphi \in \Phi$. First, $\langle \mathcal{G}_L^{\varphi}, L\rangle$ is a solution to P. For each $x \in X$ we have $\mathcal{G}_L^{\varphi}(x) = \mathcal{G}_L(\varphi, x) = I^{-1}(\varphi(x)) \in T$, so $\mathcal{G}_L^{\varphi}$ is a grade assignment from P into L. Furthermore, L is suitable for P: since by assumption Φ is truth compatible for P, for any given $x \in X$ there is some $\varphi \in \Phi$ such that $\alpha(x) = \varphi(x)$. Now, since L is suitable for Φ in P, $\alpha(x) = \varphi(x) \in V$.

To see that $\langle \mathcal{G}_L^{\varphi}, L\rangle$ is accurate, by its own standards, consider any $x \in X$. Because Φ is truth compatible for P, $\alpha(x) \in [\Phi(x) \cap I(I^{-1}(\alpha(x)))] \neq \emptyset$. We have, by definition of $\Phi(x)$, also $\varphi(x) \in [\Phi(x) \cap I(I^{-1}(\varphi(x)))] \neq \emptyset$. Since L masks Φ in P, therefore $I^{-1}(\alpha(x)) = I^{-1}(\varphi(x))$. By definition of $\mathcal{G}_L$ furthermore $I^{-1}(\varphi(x)) = \mathcal{G}_L(\varphi, x) = \mathcal{G}_L^{\varphi}(x)$. Putting these identities together within the scope of I: $\alpha(x) \in I(I^{-1}(\alpha(x)) = I(I^{-1}(\varphi(x))) = I(\mathcal{G}_L^{\varphi}(x))$. This argument is good for any $x \in X$, so $\langle \mathcal{G}_L^{\varphi}, L\rangle$ is an accurate solution to P, by its own standards. It is good for any $\varphi \in \Phi$, so $\langle \mathcal{G}_L, L\rangle$ is a reliably accurate solution to P with scope Φ.

For the *only if* part, suppose L does not mask Φ in P. Choose some $x \in X$ and $e \neq f \in T$ such that $\Phi(x) \cap I(e) \neq \emptyset$ and $\Phi(x) \cap I(f) \neq \emptyset$. Because Φ is truth compatible, without loss of generality $\alpha(x) \in I(e)$. (If $\alpha(x)$ is ”covered” by neither the chosen e nor f then, by the properties of interpretations, there has to be some other $g \in T$ such that $\alpha(x) \in I(g)$. By truth compatibility of Φ, $\alpha(x) \in \Phi(x) \cap I(g) \neq \emptyset$, so we can choose this g instead of e.) Now, choose any $v \in \Phi(x) \cap I(f)$, and any $\varphi \in \Phi$ such that $\varphi(x) = v$. We have $\mathcal{G}_L^{\varphi}(x) = I^{-1}(\varphi(x)) = I^{-1}(v) = f$. Because I is an interpretation and $e \neq f$, $I(e) \cap I(f) = \emptyset$, so because $\alpha(x) \in I(e)$, $\alpha(x) \notin I(f) = I(\mathcal{G}_L^{\varphi}(x))$. So $\langle \mathcal{G}_L^{\varphi}, L\rangle$ is not an accurate solution to P, by its own standards, and $\langle \mathcal{G}_L, L\rangle$ is not a reliably accurate solution to P with scope Φ. □

Theorem 8

Let $\langle \mathcal{G}_1, L_1\rangle$, …, $\langle \mathcal{G}_n, L_n\rangle$ be reliably accurate individual solutions to P with respective scopes Φ_1, …, Φ_n, where L_1, …, L_n measure some common dimension. Then

$$\langle \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_n\rangle), \quad \circ(\langle L_1, \ldots L_n\rangle)\rangle$$

is a reliably accurate solution to P with scope $\Phi_1 \times \ldots \times \Phi_n$.

Definitions. Let $L = \langle T, \succeq, I\rangle$ and $P = \langle X, \alpha\rangle$, and let Φ be a scope for problem P.

$\mathcal{G}$ is a *signaled grade assignment from* Φ *and* P *into* L if

- $\mathcal{G}$ maps each $(\vec{\varphi}, x) \in \Phi \times X$ to a grade in L (that is, $\mathcal{G}(\vec{\varphi}, x) \in T$), and
- $\mathcal{G}(\vec{\varphi}, x) = \mathcal{G}(\vec{\varphi}, y)$ whenever for each component φ_i of $\vec{\varphi}$, $\varphi_i(x) = \varphi_i(y)$.

Replace $\vec{\varphi}$ by φ for an *individual* signaled grade assignment.

$\langle \mathcal{G}, L\rangle$ is an *(individual) signaled solution* to P *with scope* Φ if $\mathcal{G}$ is an (individual) signaled grade assignment from Φ and P into L.

A signaled solution $\langle \mathcal{G}, L\rangle$ to P with scope Φ is a *reliably accurate solution to* P with scope Φ if for every $\vec{\varphi} \in \Phi$, $\langle \mathcal{G}^{\vec{\varphi}}, L\rangle$ is an accurate solution to P, by its own standards. Replace $\vec{\varphi}$ by φ for a reliably accurate *individual* solution.

Let $\langle \mathcal{G}_1, L_1\rangle$, …, $\langle \mathcal{G}_n, L_n\rangle$ be reliably accurate individual solutions to P with respective scopes Φ_1, …, Φ_n. A function $\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_n\rangle)$ is defined by recursion that maps each pair $(\vec{\varphi}, x)$, with $\vec{\varphi} \in \Phi_1 \times \ldots \times \Phi_n$ and $x \in X$, to a grade in $\circ(\langle L_1, \ldots, L_n\rangle)$. Where $\varphi_i \in \Phi_i$ and $x \in X$, put

- $\circ(\langle \mathcal{G}_1\rangle)(\langle \varphi_1\rangle, x) = \mathcal{G}_1(\varphi_1, x)$, and
- $\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1}\rangle)(\langle \varphi_1, \ldots, \varphi_{m+1}\rangle, x) =$
  $\langle \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m\rangle)(\langle \varphi_1, \ldots, \varphi_m\rangle, x), \quad \mathcal{G}_{m+1}(\varphi_{m+1}, x)\rangle$.

Proof of theorem 8. An induction on n. The base case is trivial, since $\circ(\langle \mathcal{G}_1\rangle) = \mathcal{G}_1$ (modulo identifying any singleton $\langle \varphi\rangle$ with φ), and $\circ(\langle L_1\rangle) = L_1$. The induction step is more involved. There are two main things to be shown: (1) $\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1}\rangle)$ is a signaled grade assignment from $\Phi_1 \times \ldots \times \Phi_{m+1}$ and $P = \langle X, \alpha\rangle$ into $\circ(\langle L_1, \ldots, L_{m+1}\rangle)$; and (2) for any $\langle \varphi_1, \ldots, \varphi_{m+1}\rangle \in \Phi_1 \times \ldots \times \Phi_{m+1}$,

$$\langle \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1}\rangle)^{\langle \varphi_1,\ldots,\varphi_{m+1}\rangle}, \quad \circ(\langle L_1, \ldots L_{m+1}\rangle)\rangle$$

is an accurate solution to P, by its own standards.

(1) There are again two requirements. The first is that (i) for any given $\langle \varphi_1, \ldots, \varphi_{m+1}\rangle \in \Phi_1 \times \ldots \times \Phi_{m+1}$, and for any $x \in X$,

$$\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1}\rangle)(\langle \varphi_1, \ldots, \varphi_{m+1}\rangle, x)$$

is a grade in $\circ(\langle L_1, \ldots, L_{m+1}\rangle)$. The second is that (ii) for any $x, y \in X$,

$$\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1}\rangle)(\langle \varphi_1, \ldots, \varphi_{m+1}\rangle, x) = \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1}\rangle)(\langle \varphi_1, \ldots, \varphi_{m+1}\rangle, y)$$

if for each $1 \leq i \leq m+1$, $\varphi_i(x) = \varphi_i(y)$. We verify (i) and (ii) in turn.

(i) Consider any particular such $\langle \varphi_1, \ldots, \varphi_{m+1}\rangle$ and x. Since $\mathcal{G}_{m+1}(\varphi_{m+1}, x)$ is a grade in L_{m+1}, by the definition of $\circ(\langle L_1, \ldots, L_{m+1}\rangle)$ it is sufficient that (a)

$\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m\rangle)(\langle \varphi_1, \ldots, \varphi_m\rangle, x)$ is a grade in $\circ(\langle L_1, \ldots L_m\rangle)$
and, letting $I_{1,\ldots,m}$ and I_{m+1} be the interpretation functions of $\circ(\langle L_1, \ldots L_m\rangle)$ and L_{m+1}, that (b)

$$\begin{gathered} I_{1,\ldots,m}(\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m\rangle)(\langle \varphi_1, \ldots, \varphi_m\rangle, x)) \\ \cap \\ I_{m+1}(\mathcal{G}_{m+1}(\varphi_{m+1}, x)) \end{gathered}$$

is non-empty.

By induction hypothesis, $\langle \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m\rangle), \quad \circ(\langle L_1, \ldots L_m\rangle)\rangle$ is a reliably accurate solution to P with scope $\Phi_1 \times \ldots \times \Phi_m$. It follows immediately that (a) $\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m\rangle)(\langle \varphi_1, \ldots, \varphi_m\rangle, x)$ is a grade in $\circ(\langle L_1, \ldots L_m\rangle)$. Furthermore,

$$\langle \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m\rangle)^{\langle \varphi_1,\ldots,\varphi_m\rangle}, \quad \circ(\langle L_1, \ldots L_m\rangle)\rangle$$

is an accurate solution to P, by its own standards, and so by the assumptions of the theorem is $\langle \mathcal{G}_{m+1}^{\varphi_{m+1}}, \quad L_{m+1}\rangle$. This secures (b), since

$$\alpha(x) \in I_{1,\ldots,m}(\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m\rangle)^{\langle \varphi_1,\ldots,\varphi_m\rangle}(x))$$

and

$$\alpha(x) \in I_{m+1}(\mathcal{G}_{m+1}^{\varphi_{m+1}}(x)).$$

(ii) Suppose for given $x, y \in X$ that for each $1 \leq i \leq m+1$, $\varphi_i(x) = \varphi_i(y)$. By induction hypothesis $\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m \rangle)$ is a signaled grade assignment from $\Phi_1 \times \ldots \times \Phi_m$ and P into $\circ(\langle L_1, \ldots, L_m \rangle)$, and by the assumptions of the theorem $\mathcal{G}_{m+1}$ is a signaled grade assignment from Φ_{m+1} and P into L_{m+1}. Therefore $\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m \rangle)(\langle \varphi_1, \ldots, \varphi_m \rangle, x) = \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m \rangle)(\langle \varphi_1, \ldots, \varphi_m \rangle, y)$ and $\mathcal{G}_{m+1}(\varphi_{m+1}, x) = \mathcal{G}_{m+1}(\varphi_{m+1}, y)$. Thus we have as required:

$$\begin{gathered}
\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1} \rangle)(\langle \varphi_1, \ldots, \varphi_{m+1} \rangle, x) \\
= \\
\langle \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m \rangle)(\langle \varphi_1, \ldots, \varphi_m \rangle, x), \quad \mathcal{G}_{m+1}(\varphi_{m+1}, x) \rangle \\
= \\
\langle \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m \rangle)(\langle \varphi_1, \ldots, \varphi_m \rangle, y), \quad \mathcal{G}_{m+1}(\varphi_{m+1}, y) \rangle \\
= \\
\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1} \rangle)(\langle \varphi_1, \ldots, \varphi_{m+1} \rangle, y)
\end{gathered}$$

(2) Take any such $\langle \varphi_1, \ldots, \varphi_{m+1} \rangle$, and any $x \in X$. Letting $I_{1,\ldots,m+1}$ be the interpretation function of $\circ(\langle L_1, \ldots, L_{m+1} \rangle)$, to be shown is that $\alpha(x) \in I_{1,\ldots,m+1}(\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1} \rangle)^{\langle \varphi_1, \ldots, \varphi_{m+1} \rangle}(x))$. Reasoning as in (1) (i) (b) above, by induction hypothesis and assumptions of the theorem $\alpha(x) \in$

$$I_{1,\ldots,m}(\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m \rangle)^{\langle \varphi_1, \ldots, \varphi_m \rangle}(x)) \cap I_{m+1}(\mathcal{G}_{m+1}^{\varphi_{m+1}}(x)),$$

which can be rewritten

$$I_{1,\ldots,m}(\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m \rangle)(\langle \varphi_1, \ldots, \varphi_m \rangle, x)) \cap I_{m+1}(\mathcal{G}_{m+1}(\varphi_{m+1}, x)).$$

By definition of $I_{1,\ldots,m+1}$ this is equal to

$$I_{1,\ldots,m+1}(\langle \circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_m \rangle)(\langle \varphi_1, \ldots, \varphi_m \rangle, x), \quad \mathcal{G}_{m+1}(\varphi_{m+1}, x) \rangle),$$

which by definition of $\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1} \rangle)$ is just

$$I_{1,\ldots,m+1}(\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1} \rangle)(\langle \varphi_1, \ldots, \varphi_{m+1} \rangle, x)).$$

Finally, this can be rewritten

$$I_{1,\ldots,m+1}(\circ(\langle \mathcal{G}_1, \ldots, \mathcal{G}_{m+1} \rangle)^{\langle \varphi_1, \ldots, \varphi_{m+1} \rangle}(x)).$$

This completes the proof of (2) and of theorem 8. □

Theorem 9

Let $\langle \mathcal{G}, L \rangle$ be a reliably accurate solution to $\langle X, \alpha \rangle$ with scope $\Phi = \Phi_1 \times \ldots \times \Phi_n$. Let $x, y \in X$ be such that $\Phi_i(x) \cap \Phi_i(y) \neq \emptyset$, for each $1 \leq i \leq n$. Let x and y be compatible in each Φ_i. Then for each $\vec{\varphi} \in \Phi$, $\mathcal{G}(\vec{\varphi}, x) = \mathcal{G}(\vec{\varphi}, y)$.

Comment. From the proof it is clear that with $n = 1$ we have, replacing $\vec{\varphi}$ by φ, the corresponding theorem for reliably accurate individual solutions as a special case.

Definition. x and y are *compatible in* individual scope Φ_i if $\forall \varphi, \psi \in \Phi_i$ $\exists \chi \in \Phi_i$ such that $\chi(x) = \varphi(x)$ and $\chi(y) = \psi(y)$.

Proof of theorem 9. Consider any reliably accurate solution $\langle \mathcal{G}, L \rangle$ to $\langle X, \alpha \rangle$ with scope $\Phi = \Phi_1 \times \ldots \times \Phi_n$. Take any i, $1 \leq i \leq n$. Under the assumptions of the theorem, choose $v \in \Phi_i(x) \cap \Phi_i(y)$. There are $\varphi_i, \psi_i \in \Phi_i$ such that $\varphi_i(x) = v = \psi_i(y)$. Because x and y are compatible in Φ_i there is $\chi_i \in \Phi_i$ such that $\chi_i(x) = \chi_i(y)$. Do this n times to obtain $\vec{\chi} = \langle \chi_1, \ldots \chi_n \rangle \in \Phi = \Phi_1 \times \ldots \times \Phi_n$ such that $\forall i$, $\chi_i(x) = \chi_i(y)$. Because $\mathcal{G}$ is a signaled grade assignment, $\mathcal{G}(\vec{\chi}, x) = \mathcal{G}(\vec{\chi}, y)$. Since $\langle \mathcal{G}, L \rangle$ is reliably accurate it is, as shown in the paper, reliable: for every $\vec{\varphi}, \vec{\psi} \in \Phi$ and every $z \in X$, $\mathcal{G}(\vec{\varphi}, z) = \mathcal{G}(\vec{\psi}, z)$. For every $\vec{\varphi} \in \Phi$ therefore $\mathcal{G}(\vec{\varphi}, x) = \mathcal{G}(\vec{\chi}, x) = \mathcal{G}(\vec{\chi}, y) = \mathcal{G}(\vec{\varphi}, y)$. $\square$