WHY ARE THE HUMAN SCIENCES HARD? TWO NEW HYPOTHESES

[BLINDED]

ABSTRACT. We present two novel hypotheses for why the human sciences are hard: (1) we are pre-committed to a very specific domain of prediction tasks in the human sciences, which limits a powerful strategy for making scientific progress—changing the prediction task to something more tractable; and (2) due to evolutionary pressures, human baseline performance is relatively high for many of the 'low-hanging fruit' prediction tasks concerning human behavior, making progress beyond this baseline challenging. We provide a formal framework for reasoning about the difficulty of disciplines and the impressiveness of their achievements in terms of their success at particular prediction tasks.

Because they deal with systems that are highly complex, adaptive and not rigorously rule-bound, the human sciences are among the most difficult of disciplines, both methodologically and intellectually...
—Drezner (2012) "A Different Agenda," *Nature*, p. 271.

Hard-to-do science is what the social scientists do and, in particular, it is what we educational researchers do. In my estimation, we have the hardest-to-do science of them all!
— Berliner (2002) "Educational research: The hardest science of all," *Education Research*, p. 18.

The natural sciences, now so fruitful in predictions, were well established by the 17th century. The social sciences are at least three centuries behind them, consequently, in the race for achievement.
—Chambliss (1958), "Prediction in Social Sciences," *Social Science*, p. 91.

To give a physical-analogy of what passes for mathematical sociology, we would have to put into a mathematical formula statements like 'if you bang them hard enough, most things crack up'. Actually this is being charitable to Simon because the last statement, though exceedingly vague, is at least true. A better physical equivalent of Simon's formalization of Homans' theories would be a sentence like 'the wind bloweth where it listeth'; which could also be written in the mathematical symbolism of vector calculus, $v_b - b_l$.
—Andreski (1974) "The human sciences as Sorcery," p. 129

1. Introduction

The human sciences, it is claimed, are hard. Propounding why they are hard has been a perennial pastime of social scientists and philosophers alike.[1] We attempt to clarify the sense in which the human sciences may (and may not) be hard, and present two novel hypotheses for why this might be so.

Our hypotheses for why the human sciences are hard are that: (1) in the human sciences we are pre-committed to a very specific domain of prediction tasks, and this limits a powerful strategy for making scientific progress, specifically, changing the prediction task to a more tractable one; and (2) due to evolutionary pressures, humans are actually already quite good at predicting the behavior of other humans relative to other prediction tasks, and so progress beyond this higher baseline performance is more challenging than in other domains in which humans exhibit a lower baseline performance.

In order to clearly formulate our hypotheses, we introduce a formal model of the 'difficulty' of disciplines and the 'impressiveness' of their achievements in which these claims can be made precise.

Our aim is to extend the set of plausible explanations for why the human sciences seem to be hard, to provide a formal framework to begin to reason about the difficulty of disciplines, and to begin a discussion regarding the sorts of evidence that might be required to adjudicate between potential explanations.

We proceed as follows. In §2, we survey extant explanations for why the human sciences are hard. In §3, we clarify what it might mean to claim that the human sciences are 'hard' and select one sensible precisification of the claim to consider: in terms of their performance on particular prediction tasks. In §4, we present a formal model for reasoning about the difficulty of disciplines in terms of their most successful predictions. We further introduce a psychologically primitive notion of 'impressiveness' of disciplines that can be conceived as distinct from but interacting with hardness. In §§5-6, we use this model to present our alternative explanations and discuss their implications. In §7, we take stock of where we are.

2. Extant Explanations

Perhaps it is Meehl in his seminal (1978) paper, "Theoretical Risks and Tabular Asterisks," who provides the most extensive catalog of reasons why the human sciences are hard. Examination of the literature reveals that most putative explanations for the difficulty of the human sciences can be thought of as falling into one of three broad categories: problems in subject matter, problems in methods, and problems in incentives.

[1]For a range of examples, see Smith (1927), Chambliss (1958), Machlup (1961), Andreski (1974), Myrdal (1972), Meehl (1978), Hedges (1987), Kuhn (1991), Finkelstein (2005), Stern and Feller (2007) Drezner (2012), Stephan (2012), Muthukrishna and Henrich (2019), and Eronen and Bringmann (2021).

The most common type of explanation locates the difficulty of the human sciences in its subject matter: human behavior is somehow inherently more challenging than that of other sciences.

As examples, we find: Drezner (2012), who locates the difficulty of the social sciences in the complex, dynamic nature of human behavior; Finkelstein (2005), who posits that human action, perception, feeling, and decisions are not describable by 'systems of invariant relations'; Machlup (1961), who emphasizes that social phenomena are more complicated, less amenable to universal generalizations, and harder to measure; Eronen and Bringmann (2021), who emphasize the lack of robust phenomena with which to ground theories, produce valid constructs, and discern causal relationships; and Meehl (1978), who rounds out the list by adding the confounders of individual differences, polygenic heredity, divergent causality, lack of law-likeness, context-dependence, the importance of cultural factors, unknown critical events, feedback loops, autocatalytic processes, randomness, and the 'sheer number of variables' endemic to the human sciences.

A second type of explanation posits that problems in methods are what impede the human sciences. For example, poor statistics education or the difficulty of conducting rigorous experiments to test many hypotheses of interest.

Among those who posit problems with methods as a primary cause, we can find: Andreski (1974), who locates the problem in the conspiracy of poor methods; Muthukrishna and Henrich (2019), who focus on the lack of a unifying theoretical framework; each Smaldino (2019), van Rooij and Baggio (2020), Guest and Martin (2021), and Borsboom et al. (2021) who attest to the need for formal methods; Hedges (1987), who posits an empirically less cumulative character of inquiry; Myrdal (1972), who notes the vulnerability to bias involved in reasoning about topics rife with moral and political implications; and Meehl (1978), who adds ethical constraints on research, a lack of consistency tests, a lack of operational definitions, problems with the meaningfulness of units of measurements and psychometric scales (1967), and the mess of confused statistical practices surrounding null hypothesis significance testing (1990).

The last, less-common sort of explanation looks to problems of incentives that may be particularly acute in the human sciences. Here, we find: Stern and Feller (2007), who posit that a lack of funding draws fewer scholars to the human sciences and slows progress; and Andreski (1974) and Stephan (2012), who conjecture that, in the absence of clear verifiability of many claims, social dynamics and perverse incentives hold greater sway, deranging inquiry.

A unifying feature of such accounts is that they attempt to explain why the human sciences are hard by focusing on particular, local facts that make the human sciences hard—those about human behavior, current methods, and institutional incentives.

The hypotheses we proffer, on the other hand, focus on certain more general, structural features of the human sciences. Just as one might explain the success of a scientific theory by pointing to specific features of that theory, or by pointing to general features of the process that produced it (Van Fraassen 1980, p. 40), we will focus on such general features.

The explanations we present here are not mutually exclusive with the explanations that others have offered; they can all be true at the same time.[2] Furthermore, we do not claim that our two explanations explain the entire success gap, nor do we claim that the explanations we offer are even the main contributors. Rather, we think that our two explanations are plausible contributors to the perceived difficulty gap between the human sciences, and other sciences.

3. Sharpening the Claim

To a first approximation, our subject of interest can be articulated as follows:

SSH1. *The human sciences are hard.*

But, hard relative to what? Presumably, the comparison is with other scientific domains, such as the physical, chemical, and biological sciences. Let us restate the thesis with the comparison made explicit.

SSH2. *The human sciences are hard relative to other sciences.*

Now, we need to ask: hard in what respect? One interpretation is that the human sciences are hard in the sense that they require greater effort or intellect to engage in, relative to other sciences. This seems false; and not what any of the authors we considered claim.[3] A more natural interpretation is that the human sciences are hard in that they produce less predictive and explanatory success relative to effort than other sciences. Indeed, this is sometimes stated explicitly.[4]

But how do we assess predictive and explanatory success? This is not easy; there is no standard, well-defined measure of these concepts for a science as a whole. We chose to focus on performance in specific prediction tasks and operationalize the assertion as follows:

SSH3. *In the human sciences, we tend to exhibit worse performance in prediction tasks of interest relative to the performance in prediction tasks of other sciences.*

[2]Even so, having novel hypotheses on the table should change our estimates of the relative contribution of the preexisting explanations to the difficulty of the human sciences.

[3]Thus we are not considering the type of difficulty to which someone might be referring when they say, "physics is a difficult subject because it involves a lot of sophisticated mathematics"—we are *not* considering how difficult it is for a practitioner of a certain discipline to practice.

[4]Cf. Tetlock (2005), Yarkoni and Westfall (2017), Hofman et al. (2017), and The Forecasting Collaborative et al. (2023).

This operationalization expresses the comparative nature of the claim, while also directing our focus toward prediction tasks, which serves as a proxy for assessing the difficulty of scientific inquiry across disciplines.

The choice to focus on prediction tasks is for two primary reasons. First, prediction inherently encapsulates information regarding other forms of scientific achievement; identifying causal mechanisms or providing accurate explanations can be viewed as specialized forms of prediction tasks (Yarkoni and Westfall 2017).

Second, alternative markers of scientific achievement, such as explanation and understanding, are often ambiguous and subject to debate (Woodward et al. 2017; De Regt 2017). In contrast, we have access to precise, well-understood, and largely uncontested definitions of prediction in terms of minimizing inaccuracies under proper loss functions (Gneiting and Raftery 2007).[5] Given these considerations, performance in prediction tasks serves as a reasonable first-pass surrogate for overall scientific proficiency.

Consider some characteristic prediction tasks within each domain. In the human sciences, prediction tasks might include forecasting economic trends, predicting voting behavior, or estimating the occurrence of social phenomena. In contrast, prediction tasks in the physical, chemical, and biological sciences might encompass forecasting weather patterns, predicting chemical reactions, or projecting trajectories of physical entities.

Furthermore, notice that we also specify that it is prediction tasks *of interest* on which we focus. This underscores the relevance of restrictions on the set of tasks relating to the subject matter. The impact of restricting the human sciences to a very specific subset of possible questions regarding human behaviours is an important part of the explanations we offer. That this happens in science is well-studied. Indeed, it is a core insight from the values in science literature, that a variety of values—from conceptual simplicity to societal impact—come into play at various stages of scientific inquiry, and certainly in our choice of which research questions to formulate and pursue (Kuhn 1981; Douglas 2000; Longino 2004).

It is important to clarify that we do *not* assert that the sole objective of science is to excel in prediction tasks, even when when conceiving of the set of prediction tasks as subsuming predictions regarding causation and explanation. However, in grappling with the notion of difficulty, one must begin somewhere. We think that focusing on prediction tasks as the unit of performance can provide valuable insights into the relative challenges encountered across scientific disciplines. And so, it is where we start.

[5]Despite their differences, both frequentist and Bayesian accounts share common ground regarding the definitions of concepts such as prediction, accuracy, precision, and statistical consistency.

4. The Difficulty of Disciplines: A Model

In order to characterize different hypotheses about what makes the human sciences, or any science, hard, we introduce a simple model. The model aims to capture certain qualitative features of the structure of inquiry; it can be modified as necessary to incorporate more subtle structure.

We imagine inquiry as a set of prediction tasks, $\mathbb{T}$. $\mathbb{T}$ is just a set. If one wanted, one could add more structure to $\mathbb{T}$ or its members. Instead of doing so directly, we introduce a function, $H : \mathbb{T} \to \mathbb{R}^+$. $H(t)$ represents the hardness of task $t \in \mathbb{T}$.

For example, a member t of $\mathbb{T}$ could be predicting the trajectory of a comet as it approaches Earth, whether or not a drug will reduce a symptom in a patient, or predicting the outcome of an election in Canada.

A *discipline* D is a subset of $\mathbb{T}$. For example, the prediction task "predict the trajectory of this comet" belongs to the discipline physics, but not to the discipline biology. Disciplines can also overlap; the prediction task "predict the outcome of this election" might belong to both political science and social science.

If $\mathbb{T}$ is finite, cardinality is a fine notion of size of disciplines.[6] For example, we can say that one discipline D_1 is larger than another D_2 if $|D_1| > |D_2|$. Note that size here is about the set of problems in the purview of the discipline. Sometimes it might also be the case than one discipline is a strict subset of the other: $D_1 \subset D_2$. For example, organic chemistry is a subdiscipline of chemistry.

We imagine that a scientist attached to a certain discipline is tasked with solving tasks in that discipline. Thus, the discipline restricts what practitioners of that discipline can investigate. In the real world, the way in which prediction tasks are restricted would correspond to the social constraints of inquiry: it would be difficult for a philosopher to secure a large grant to investigate a question in synthetic chemistry.

This is the basic set-up. $\mathbb{T}$ describes the tasks; H their difficulty. Of course, we are *uncertain* about the difficulty of various tasks. In order to model this uncertainty, we introduce a sample space. We assume here that $\mathbb{T}$ is known; all of the uncertainty is about the difficulty of tasks. That is, it is about H.

Our sample space $\mathcal{S}$ is a subset of all possible functions from $\mathbb{T}$ to the positive reals: $\mathcal{S} \subset \mathbb{R}^{+\mathbb{T}}$. Our particular choice of $\mathcal{S}$ depends on how we conceptualize tasks and their difficulties. For example, if each task is a binary prediction task, and difficulty is measured by the expected Brier score, then it makes sense for $\mathcal{S} = [0, 1]^{\mathbb{T}}$.

[6] If T is not finite, then the subset relation still makes a good comparison notion when it applies. To compare the size of disciplines when it doesn't apply, we might use a different notion of size, such as measure. In this paper we consider the finite case for ease of exposition.

Uncertainty about a specific task $t \in \mathbb{T}$ is parasitic on our uncertainty about H. Thus, we can represent the difficulty of a task $t \in \mathbb{T}$ as a random variable, $\Delta_t : \mathcal{S} \to \mathbb{R}^+$, given by $\Delta_t(H) = H(t)$.

In this set-up, it now makes sense to say, for example, that the difficulties of tasks are IID, or other similar statements. Note that this is implicitly a statement about a suppressed probability distribution, P, over an algebra formed over $\mathcal{S}$. Moving forward, we continue to suppress these details, focusing instead on the relationships between the random variables representing the difficulty of tasks.

Following our discussion in §3, we imagine the goal of scientists as solving prediction tasks. We assume that how well a scientist is able to solve a task is a decreasing function of its difficulty.[7] Thus, without loss of generality, we will consider the difficulty of an attempted task as a measure of how poorly a task was solved.

We have a precise sense of what it is for a task to be difficult. But what about the difficulty of a whole discipline? Aggregating the difficulties of a set of tasks to yield the difficulty of a discipline involves many choices. For example, one could take the average difficulty of tasks in the discipline, or the most common difficulty in a discipline.

Here we introduce a discipline-level measure of difficulty that is meant to capture an important part of how we make *judgments* about the difficulty of disciplines. We measure the difficulty of a discipline by how poor its best performance was. Or, rather more intuitively, we measure how *easy* a discipline is by its greatest success.

To make this precise, we introduce the (slightly overloaded) notation of

$$H(D_i) := \min\{h | \exists t \in D_i : h = H(t)\}$$

to describe the *difficulty* of a discipline D_i. That is, we identify how difficult a discipline is with the easiest task in its purview.

This flat-footed approach tracks a core intuition behind the *judgment* that the human sciences are hard, and the physical sciences easier: we hold up the impressive successes of modern physics, and the less impressive successes of modern human science, as examples. We don't focus as much on how poorly each discipline fares on their worst tasks.[8] Thus, while for the remainder of the paper we will often call $H(D)$ the difficulty of discipline D, we might also think of $H(D)$ as capturing how difficult D *seems* to be, since H throws away information about how well a discipline performs in general, and instead only focuses on its most impressive successes.

[7]Once again, when we speak of difficulty, we are *not* talking about how difficult it is for a scientist to do the tasks needed to do science. For example, if successfully predicting the trajectory of a comet requires a lot of challenging formal work, but the prediction itself is very successful because comet trajectories are regular, we do not consider this a difficult task. Compare this to footnote 3.

[8]Basically every science has questions within its purview that it cannot yet answer.

We do *not* claim that this is the uniquely best measure of the difficulty of disciplines. Indeed, both of our explanations in the remainder of the paper can be understood as showing that an inference from *perceived* difficulty, as captured by H, to some other measure of difficulty, such as *average* difficulty, is not straight-forward, and can fail. Rather, we use H because we think that paying attention to the impressive successes of a discipline plays an important role in producing the judgement that disciplines are hard or easy, and we want to understand better what produced this judgement.

Finally, notice that we can also consider how difficult we *expect* a discipline to be as well, given our uncertainty about the difficulty of tasks, by taking the expected value of the difficulty of a discipline: $\mathbb{E}(H(D_i))$.

5. Rigid Demands Hypothesis (RD)

We now present out first novel hypothesis. The *rigid demands* hypothesis says that we should expect the human sciences to seem less impressive simply in virtue of the stricter demands we place on them, compared to other sciences.

For example, consider the physical sciences. In physics, it is very difficult to predict the motion of any particular particle in a gas. If we required physicists to be able to predict the motion of specific particles, and measured the success of physics based on these tasks, it would not appear very impressive. However, it is much more tractable to predict the temperature of a gas, given, for example, its volume and pressure. Judging physics by this prediction task, it seems much more successful. Indeed, choosing a more tractable problem is a core strategy for making progress in a field.

In contrast, consider education sciences. In this field, demands are much more rigid: we want to know whether this particular intervention will help, for example, increase the future earnings of graduates. This is a very specific problem, and the education researcher does not have the flexibility to work on a more tractable problem that will not be useful for policy makers.

Of course the real world is complex: many fields in physics are constrained by things like military and civilian applications, and the education researcher does have some room to adjust her inquiry. However, the freedom to redefine research questions to increase tractability does vary by field. This is the aspect of a discipline that RD uses to help explain why the human sciences seem less impressive to us.

RD *itself* does not say that the human sciences are *not* harder than the other sciences. Rather, it says that our more rigid demands *also* contribute to why the human sciences seem less impressive. Again, this is because we expect our explanations to play some role in explaining why we take the human sciences to be hard, but do not expect them to be the only contributors.

To make clear how this works, we will consider a version of RD that *does* assume that there are no other contributors. In other words, we will assume

that the task-by-task difficulty between different disciplines is equal in some way, and show that RD can still make the expected value of H different across disciplines. Call this the "strong rigid demands" hypothesis (SRD). We might state it in natural language as follows:

SRD. *The human sciences are not intrinsically harder than other sciences. Rather, the demands we place on the prediction tasks for the human sciences are stricter than the demands we place on other sciences. Thus, social scientists do not have as much freedom to investigate easy tasks. This means that we should expect their best successes to be less impressive.*

Our model allows us to make this more precise. When we say that "the human sciences are not intrinsically harder than other sciences",[9] what we mean is that the distribution of the difficulty of tasks in the human sciences is identical to that in other sciences.[10] A strong version of this would be that all tasks in $\mathbb{T}$ are independent and identically distributed (IID).[11] When we say that the demands we place on the human sciences are stricter than those we place on other sciences, what we mean is that the size of the discipline *human sciences* is smaller than the sizes of other sciences, say, for example, *physical sciences*: $|D_{hs}| < |D_{ps}|$.

The last part of the claim is an inference from the stricter demands to lower expected best success. If we consider a version of the model with the IID assumption for all tasks, we can show that this holds as a result of the difference in cardinality.

Proposition 1. *Let the set of all task difficulties, $\Delta = \{\Delta_t\}_{t\in\mathbb{T}}$ be IID and the distributions non-degenerate (i.e., the distribution does not assign probability 1 to any real number) and let $|D_1| < |D_2|$. Then $\mathbb{E}(H(D_1)) > \mathbb{E}(H(D_2))$.*

This proposition expresses the core reasoning underlying SRD.[12] Even assuming that tasks are equally difficult across disciplines, how successful we expect a discipline to be, as measured by $\mathbb{E}(H(D))$, is an increasing function of its size. A corollary immediately follows:

Corollary 1. *Let D be a discipline, and assume that the tasks in D are IID and the distributions non-degenerate. Then, as $|D|$ decreases, holding the distribution of the difficulty of tasks in D fixed, $\mathbb{E}(H(D))$ increases.*

This shows that the size of the discipline really does affect the expected difficulty of the discipline. Thus, even if disciplines do vary in expected

[9]This is of course a different notion of difficulty than that captured by $H(D)$.

[10]Really, what we wrote here is even stronger, since we posit equal distribution, and SRD only states an inequality.

[11]There are, of course, other ways to make this precise.

[12]To be very clear: this proposition, and the ones that follow, are themselves of course trivial to show. We do not take these propositions *themselves* to be the core contribution of the article; rather, we include them to show very precisely the kind of phenomenon that we suggest is at play in our jugdements that the human sciences are hard.

difficulty of their tasks, changing their sizes will also change the expected difficulties of the disciplines. This shows that rigid demands play a role in determining how successful a science should seem relative to others, even if we expect one discipline to have a different distribution of task difficulties than another.[13]

There are other features of disciplines that can affect the range of prediction tasks in their purview, beyond the demands we place on them. For example, another way that the qualitatively same outcome can be produced is for a discipline to have existed for longer and hence formulated more problems and made more attempts at solving problems. This will increase the cardinality of the set of its prediction tasks and thus the expectation for its most successful solutions. Again, in these cases, it can be that one discipline *seems* harder than another, even though the underlying distribution of prediction task difficulty is the same.

6. Fruit in the Hand Hypothesis (FTH)

Our second novel hypothesis is the *fruit in the hand* hypothesis. It says that, due to our evolutionary endowment and distinctive enculturation, many of the easy prediction tasks in the social domain have already been solved; we have already picked all or most of the low hanging fruit. But we do not count these successes are part of the successes of the human sciences. This leaves the harder problems, high up in the tree, for the social scientists to tackle. Furthermore, our evolutionary and cultural inheritance has not given us the same gifts when it comes to the other sciences.

For example, one can predict very accurately that, if one were in a room full of philosophers quietly writing articles, and one suddenly and without warning screamed and threw a glass of water at the wall, shattering it, people would stop writing their papers and react with surprise and fear.[14] We do not include such prediction tasks as part of the human sciences, even though they concern the behaviour of humans in response to stimuli.

Similar to the rigid demands hypothesis, we can write down a *strong* version of the fruit in the hand hypothesis, in order to clarify how the weak version works.

---

[13]Notice that, from a Bayesian perspective, learning facts about the cardinality of the set of predictions tasks in a discipline will typically affect our expectations about the relative contribution of inherent task difficulty across domains to the observed difference of success between disciplines. We leave the development of such a Bayesian model to future work.

[14]One author did briefly consider testing this hypothesis as he wrote these words in such a room; he did not yield to the temptation. Such is the confidence with which this prediction can be made. We trust that the reader also does not need to test this hypothesis, or a very similar one, to have high confidence that this is true. And that is precisely the point; we exclude such prediction tasks from the sciences, *because* they are, in a sense, already solved.

SFTH. *Human prediction tasks are not intrinsically harder than prediction tasks concerning other types of objects. Rather, due to evolutionary pressures, humans are already good at prediction in a broad range of tractable prediction tasks related to human behaviour. Thus, these tasks are not considered part of the human sciences. There was not as much pressure in other domains, and so humans did not already solve as many prediction tasks. This means that we should expect the best successes of the human sciences to seem less impressive than the other sciences.*

There are two ways to express this claim in the model. The first involves starting with a discipline, and then removing easy prediction tasks from it. The second is more involved, and posits a correlation between difficulty and a new property of prediction tasks, *impressiveness*. We consider each precisifiction in turn.

6.1. **Plucking the Fruit.** One way of making SFTH precise is to imagine starting with a discipline, and then removing some of the easiest prediction tasks from it. As the easiest tasks are removed, the expected difficulty of the discipline goes up. This reasoning is witnessed by the following observation.

Proposition 2. *Let D be a discipline, and assume that the tasks in D are IID and the distributions non-degenerate. Let $\ell \in \arg\min_{t \in D} h(t)$, and let t be a task in D chosen uniformly at random. Then $\mathbb{E}(H(D - \{\ell\})) \geq \mathbb{E}(H(D - \{t\}))$ and $\mathbb{E}(H(D - \{\ell\})) \geq \mathbb{E}(H(D))$.*

By comparing what would happen if we removed the easiest task to what would happen if we removed a random task, we see that the effect of removing the easiest task goes beyond the observation in §5.[15]

Again, consider the case of screaming and throwing a glass of water at the wall in a room full of quietly working philosophers. This is an easy prediction task; like the person throwing the glass, the solution screams out in a way that is difficult to ignore. It is also one that we do not consider a finding of the human sciences, even though it concerns human behaviour. Removing this task from the domain of the human sciences increases the expected difficulty of the discipline.

Once such easy prediction tasks are removed from the human sciences, what is left over are more difficult problems. These often involve more subtle patterns of behaviour, which one must use sophisticated statistics to even try to identify.

In contrast, many scientific disciplines that we consider very successful explore a space of possible research questions partially populated by patterns that speak clearly, since such patterns weren't already incorporated into our intuitions and native prediction capabilities by evolutionary processes, and thus excluded from the discipline. Scientists in such disciplines commonly

[15]The uniformly random assumption is not actually needed; the proof in the appendix shows that this holds for any task one removes. However, the idea of removing a random task is intuitive, which is why we stated it as such in the main text.

dismiss equivocal test results as unsatisfactory and opt to pursue alternative lines of inquiry that promise unequivocal findings. Within such domains, the strength of many inferences obviates the necessity for statistical tools. Indeed, as articulated by biologist Pamela Reinagel, prevailing sentiments within such fields assert that "If you needed to do a statistical test, you just did a bad experiment," and "If you needed statistics, you are studying something so trifling it doesn't matter" (Reinagel 2022).

6.2. **Tell Me Some Fruit I *Don't* Know.** The second way of making SFTH precise is to introduce an additional property of prediction tasks, *impressiveness*. Impressiveness is a measure of how impressed we would be if science solved a certain prediction task.

Impressiveness allows us to express SFTH in a more graded way. Instead of removing a prediction task from the domain of a science like in §6.1, with impressiveness we can keep the task in the domain of the science, and just decrease the impressiveness of that task.

To capture the difference in how impressed we are by solving tasks in different disciplines, we have a different impressiveness function for each discipline. Formally, we introduce impressiveness as a function $I_D : \mathbb{R}^+ \to \mathbb{R}^+$. The interpretation of $I_D(x) = y$ is that, when x is the difficulty of some prediction task in D, then y is the impressiveness of that task.[16]

We can then combine I_D with a prediction task in order to define an *unimpressiveness weighted difficulty*. We define it as follows:

$$\mathbb{U}(t) := \frac{H(t)}{I_D(H(t))}$$

whenever $t \in D$.

The idea is that, even if we *include* easy tasks from the human sciences (instead of excluding them from the discipline like in §6.1), we might not be very impressed by the science that solves them. If I already have a bunch of apples on hand, and you tell me about another, I might rightfully reply, *tell me some fruit I don't know.*

We capture this by dividing the difficulty of a task by its impressiveness. Notice that, if $I_D(x) = 1$, then there is no change. However, if $I_D(x) < 1$, then $\mathbb{U}_D(x)$ is larger. Given that in the model smaller numbers indicate more successful predictions, this captures the reasoning that, if a task is less impressive, it having been solved doesn't make the science appear as successful as if it were impressive.

Recall the example of throwing a glass against the wall to startle people and predicting what will happen (task w). Even though one could very

[16]This is merely one way of introducing impressiveness in the model, and a rather inflexible one at that. For example, impressiveness has to be a function of a task's difficulty, and thus ignores other features of the task. In a richer model one could of course have a more sophisticated way of expressing impressiveness that is more general. Here, however, given that we are interested in looking at the relationship between difficulty and impressiveness, the current formalization is sufficient.

successfully solve this task, it isn't very impressive, due to our evolutionary and cultural inheritance. Thus, even if we included it in a scientific domain, it is so unimpressive that $I_{hs}(H(w))$ would be low, and thus $\mathbb{U}_{hs}(w)$ would be high.

Taking unimpressiveness weighted difficulty as our measure of success instead of H as in the base model now allows us to express statements about what seems impressive to us, and how that relates to difficulty in specific disciplines. As with difficulty, we define a discipline level version:

$$\mathbb{U}(D_i) := \min\{h | \exists t \in D_i : h = \mathbb{U}_{D_i}(t)\}$$

With all of this on the table, the suggestion is the following. Given that we are fine-tuned by evolution and enculturation to be better at predicting social phenomena than we are other phenomena (the specific location of mitochondria, plasma dynamics), the impressiveness function of the human sciences assigns a lower impressiveness to easy tasks than the impressiveness function of the other sciences. Thus, the expected unimpressiveness weighted difficulty of the human sciences is higher, even though the difficulty of the disciplines is the same.

Proposition 3. *Let D_1 and D_2 be disciplines. Let them have the same cardinality, and let the difficulty of prediction tasks across both disciplines be IID and non-degenerate. Let I_{D_1} and I_{D_2} be such that $I_{D_1}(x) \geq I_{D_2}(x), \forall x \in \mathbb{R}^+$. Then, $\mathbb{E}(\mathbb{U}(D_1)) \leq \mathbb{E}(\mathbb{U}(D_2))$. Furthermore, if $I_{D_1} > I_{D_2}$ with positive probability, then this inequality is strict.*[17]

This proposition shows that impressiveness can affect how difficult we find a science to be, even if the underlying distribution of difficulty of tasks is the same. Of course, even holding fixed the IID and cardinality assumptions, the conditions of the proposition are not necessary for the conclusion to follow. There are many ways that domain-relative impressiveness can change the expected unimpressiveness weighted difficulty of a discipline.

Folk psychology has subsumed a sizable subset of prediction tasks regarding human behavior. Many of these subsumed tasks are impressive in ways we tend to take for granted. For example, adequate solutions to basic problems of celestial motion are often achieved with a handful of second-order partial differential equations, while adequate solutions to basic tasks in natural language processing have only been broached by neural networks consisting of non-linear transformations of many billions of parameters. As Fodor notes in the first chapter of *Psychosemantics* (1987), psychology as a science needs to surpass common sense folk psychology to be impressive.

[17]This last bit is an abuse of notation, since $I_{D_i}, i \in \{1, 2\}$ is not a random variable, but gets the point across. We define the condition precisely in the proof.

7. Discussion

We hope to have furnished a fruitful formalism with which to think about the difficulty of disciplines, and to have presented two novel hypotheses about how it can *appear* that the human sciences, in particular, are hard. Specifically, we have showed how focusing on the most impressive successes of disciplines, and noticing that they differ, does not itself allow one to conclude that the distribution of task difficulties within a discipline differ between disciplines. We showed this as a stepping stone to the insight that, even in situations in which there are underlying differences, the mechanisms we identified will still change the expected observed difficulty of disciplines.

As we have stated many times, we think that our hypotheses only explain *part* of the judgment that the human sciences are difficult. We expect that the distributions of task difficulties do vary across disciplines, and very likely in ways proposed by others. However, we do think that our hypotheses show that inferring properties of the distributions of difficulties of tasks from observed success is a subtle business.

For this reason, we are excited to explore how the various suggestions in the literature interact with each other. The formalism we provided here should also allow us to reason about several of the extant explanations discussed, both individually and in tandem. Once formalized, we can represent candidate hypotheses together and observe how they interact to change the apparent difficulty of a science. This will allow us to get a more fine-grained understanding of the relative contribution of different hypotheses.

That said, work is required to make these hypotheses amenable to clearer evaluation. This will include identifying, empirically, the units of prediction tasks across some representative sciences, formulating meaningful measures of their difficulty,[18] and ascertaining the relative compatibility of the data with the candidate hypotheses. This we leave for future work.

Mathematical Appendix

Proof of Proposition 1. Let the set of all task difficulties, $\Delta = \{\Delta_t\}_{t\in\mathbb{T}}$ be IID and the distributions non-degenerate, and let $|D_1| < |D_2|$. We want to show that $\mathbb{E}(H(D_1)) > \mathbb{E}(H(D_2))$. Consider the case where $|D_1| = |D_2|$. It follows that $\mathbb{E}(H(D_1)) = \mathbb{E}(H(D_2))$. Now, let us add a single task t to D_2. Since hardness of a discipline is defined as the hardness of its least hard task, we know that $D_2 \cup \{t\}$ cannot be less hard than D_2. And since there is some positive probability that $H(t)$ is less than the least difficult task in D_2, then $\mathbb{E}(H(D_1)) = \mathbb{E}(H(D_2)) > \mathbb{E}(H(D_2 \cup \{t\}))$. □

Proof of Corollary 1. This follows is immediate from proposition 1. □

[18] for example, in terms of the correlation coefficients of the primary effects of studies in the published literature.

Proof of Proposition 2. Let D be a discipline, and assume that the tasks in D are IID and the distributions non-degenerate. Let $\ell \in \arg\min_{t \in D} h(t)$, and let t be a task in D chosen uniformly at random. Then there are two cases: $H(t) = H(\ell)$ and $H(t) \neq H(\ell)$.

Case 1. If $H(t) = H(\ell)$, then $H(D - \{t\}) = H(D - \{\ell\})$.
Case 2. If, however, $H(t) \neq H(\ell)$ we know that $H(t) > H(\ell)$ since $H(\ell)$ is a minimally difficult task. Furthermore, it is possible that ℓ is the unique easiest task, in which case we we have that $H(D - \{t\}) < H(D - \{\ell\})$. And since there is a positive probability that $H(t) \neq H(\ell)$, $\mathbb{E}(H(D - \{t\})) < \mathbb{E}(H(D - \{\ell\}))$.

Mixing between both cases with positive probability yields $\mathbb{E}(H(D-\{t\})) < \mathbb{E}(H(D - \{\ell\}))$. And notice that $\mathbb{E}(H(D)) < \mathbb{E}(H(D - \{\ell\}))$ follows immediately from proposition 1. □

Proof of proposition 3. Let D_1 and D_2 be disciplines. Let them have the same cardinality, and let the difficulty of prediction tasks across both disciplines be IID and non-degenerate. Let I_{D_1} and I_{D_2} be such that $I_{D_1}(x) \geq I_{D_2}(x), \forall x \in \mathbb{R}^+$.

Notice that, since $|D_1| = |D_2|$ and all the difficulties in each discipline are IID, it suffices to show that: $\mathbb{E}(\mathbb{U}(D_1)) \leq \mathbb{E}(\mathbb{U}_{D_2}(D_1))$, i.e., applying I_{D_2} to members of D_1. This is also true when we show the strict inequality.

Let the probability of the event described as "$I_{D_1} = I_{D_2}$" be given by $P(I_{D_1}(\Delta_t) = I_{D_2}(\Delta_t))$ (and similarly for $<$ instead of $=$), for any $t \in D_1$. Since the difficulties in D_1 are IID, the choice does not matter. Then, there are two cases: either $P(I_{D_1} = I_{D_2}) = 1$ or $P(I_{D_1} > I_{D_2}) > 0$.

Case 1. Let it be that $P(I_{D_1} = I_{D_2}) = 1$. Then it is clear that $\mathbb{E}(\mathbb{U}(D_1)) = \mathbb{E}(\mathbb{U}_{D_2}(D_1))$.

Case 2. Let it be that $P(I_{D_1} > I_{D_2}) > 0$. Then $\mathbb{E}(\mathbb{U}(D_1)) < \mathbb{E}(\mathbb{U}_{D_2}(D_1))$, since for each $t \in D_1$ there is positive probability that t is the uniquely minimally difficult task, and positive probability that its difficulty x lies in the range where $I_{D_1}(x) > I_{D_2}(x)$. □

References

Andreski, S. (1974) *Social Sciences as Sorcery.* Pelican Books. Penguin.

Berliner, D. C. (2002) Comment: Educational research: The hardest science of all. *Educational researcher. 31*(8), 18–20.

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A. and Haig, B. D. (2021) Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science. 16*(4), 756–766. 10.1177/1745691620969647. © The Author(s) 2021.

Chambliss, R. (1958) Prediction in the social sciences. *Social Science. 33*(2), 90–95.

De Regt, H. W. (2017) *Understanding scientific understanding.* Oxford University Press.

Douglas, H. (2000) Inductive risk and values in science. *Philosophy of Science. 67*(4), 559–579. 10.1086/392855.

Drezner, D. (2012) A different agenda. *Nature. 487*, 271.

Eronen, M. I. and Bringmann, L. F. (2021) The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science. 16*(4), 779–788. 10.1177/1745691620970586. PMID: 33513314.

Finkelstein, L. (2005) Problems of measurement in soft systems. *Measurement. 38*(4), 267–274. https://doi.org/10.1016/j.measurement.2005.09.002. The logical and philosophical aspects of measurement.

Fodor, J. A. (1987) *Psychosemantics: The problem of meaning in the philosophy of mind.* MIT Press.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association. 102*(477), 359–378. 10.1198/016214506000001437.

Guest, O. and Martin, A. E. (2021) How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science. 16*(4), 789–802. 10.1177/1745691620970585.

Hedges, L. V. (1987) How hard is hard science, how soft is soft science? the empirical cumulativeness of research. *American Psychologist. 42*(5), 443.

Hofman, J. M., Sharma, A. and Watts, D. J. (2017) Prediction and explanation in social systems. *Science. 355*(6324), 486–488. 10.1126/science.aal3856.

Kuhn, T. S. (1981) Objectivity, value judgment, and theory choice In *Review of Thomas S. Kuhn The Essential Tension: Selected Studies in Scientific Tradition and Change*, Zaret, D. (eds). Duke University Press. pp. 320–39.

Kuhn, T. S. (1991) The natural and the human sciences In *The Interpretive turn: philosophy, science, culture*, Hiley, D. R., Bohman, J., and Shusterman, R. (eds). Cornell University Press. pp. 17–24.

Longino, H. E. (2004) How values can be good for science In *Science, Values, and Objectivity*, Machamer, P. K. and Wolters, G. (eds). University of Pittsburgh Press. pp. 127–142.

Machlup, F. (1961) Are the social sciences really inferior? *Southern Economic Journal. 27*(3), 173–184.

Meehl, P. E. (1967) Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science. 34*(2), 103–115. 10.1086/288135.

Meehl, P. E. (1978) Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology (reprinted from the journal of consulting and clinical psychology, vol 46, pg 806-834, 1978). *Applied & Preventive Psychology. 46.*

Meehl, P. E. (1990) Why summaries of research on psychological theories

are often uninterpretable. *Psychological Reports*. *66*(1), 195–244.

Muthukrishna, M. and Henrich, J. (2019) A problem in theory. *Nature Human Behaviour*. *3*, 221–229. 10.1038/s41562-018-0522-1.

Myrdal, G. (1972) How scientific are the social sciences?1. *Journal of Social Issues*. *28*(4), 151–170. https://doi.org/10.1111/j.1540-4560.1972.tb00052.x.

Reinagel, P. (2022) The limited role of null hypothesis testing in biology.

Smaldino, P. E. (2019) Better methods can't make up for mediocre theory. *Nature*. *575*, 9. 10.1038/d41586-019-03350-5.

Smith, T. V. (1927) Recent developments in the social sciences. edward cary hayes. *The International Journal of Ethics*. *37*(4), 435–436. 10.1086/intejethi.37.4.2377895.

Stephan, P. (2012) Perverse incentives. *Nature*. *484*(7392), 29–31. 10.1038/484029a.

Stern, P. C. and Feller, I. (2007) *A strategy for assessing science: Behavioral and social research on aging*.

Tetlock, P. (2005) *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press. rev - revised edition.

The Forecasting Collaborative, Grossmann, I., Rotella, A., Hutcherson, C., Sharpinskyi, K., Varnum, M., Achter, S., Dhami, M., Guo, X., Kara-Yakoubian, M., Mandel, D., Raes, L., Tay, L., Vie, A., Wagner, L., Adamkovic, M., Arami, A., Arriaga, P., Bandara, K., Baník, G., Bartoš, F., Baskin, E., Bergmeir, C., Białek, M., Børsting, C., Browne, D., Caruso, E., Chen, R., Chie, B., Chopik, W., Collins, R., Cong, C., Conway, L., Davis, M., Day, M., Dhaliwal, N., Durham, J., Dziekan, M., Elbaek, C., Shuman, E., Fabrykant, M., Firat, M., Fong, G., Frimer, J., Gallegos, J., Goldberg, S., Gollwitzer, A., Goyal, J., Graf-Vlachy, L., Gronlund, S. and Hafenbrädl, S. (2023) Insights into the accuracy of social scientists' forecasts of societal change. *Nature Human Behaviour*. *7*(4), 484–501. 10.1038/s41562-022-01517-1.

Van Fraassen, B. (1980) *The Scientific Image*. Oxford University Press. New York.

van Rooij, I. and Baggio, G. (2020) Theory development requires an epistemological sea change. *Psychological Inquiry*. *31*(4), 321–325. 10.1080/1047840X.2020.1853477.

Woodward, J., Zalta, E. N. *et al.* (2017) Scientific explanation. *The Stanford*.

Yarkoni, T. and Westfall, J. (2017) Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*. *12*(6), 1100–1122.