Value Illusions in Grading

Michael Morreau

UIT—The Arctic University of Norway

michael.morreau@uit.no

Abstract

When people get together to grade things of one sort or another, there will often be some uncertainty about how they interpret applicable grades. We will see that too much of this uncertainty can make grading procedures, in a precise sense, unsound. Evaluations arrived at by collective grading are in this case not borne out by any procedures for aggregating the merit expressed by grades. They are unwarranted value illusions.
1. Value Illusions

The Arts and Humanities Research Council in Britain evaluates research proposals by convening a panel. It assigns to each proposal under consideration two “introducers,” who determine overall grades on a scale from 2 to 6. Later, after discussion, the panel agrees on a final grade for each proposal, and ranks the different ones from better to worse. This is an example of collective grading. There are many others. Students go online to evaluate courses as exceptional, good, adequate, somewhat inadequate, or unsatisfactory. In the United States, the National Rifle Association assigns politicians grades from A to a menacing F according to their support for its positions. Linguists express judgments of the grammaticality of sentences by assigning to them symbols such as ?, *, **. The Michelin guide assigns restaurants from 0 to 3 stars on the basis of evaluations by inspectors. In medical decision making, evidence is judged to be of high, moderate or low quality.

Grades are linguistic terms. And, like other expressions of natural languages, they are contextual. Just how good or bad something has to be in order to receive one grade or the other can vary from occasion to occasion and from person to person. There are good reasons for this. For one thing, different standards are applicable in different situations. What counts as good evidence in everyday life might not hold up in a legal or medical setting. Also, with some scope to shift the boundaries between grades, we often can avoid the kind of distortion that occurs when two items receive different grades even though they differ only slightly in merit. Contextuality contributes in these ways to the effectiveness of grades as vehicles of information about value.

With contextuality, though, comes a potential for equivocation. Suppose two graders both report that a body of evidence has high quality, but there is a disagreement about another: one says high, the other low. It might appear that the first body of evidence is better—or, at least, that this follows from the report. After all, its grades dominate. But, depending on the case, it doesn’t follow in the least. If it so happens that the second grader applies a higher standard, counting as low quality some evidence that the other counts as high, the correct conclusion might even be that the second body of evidence is the better one. While we must reckon with the possibility of this verbal disagreement, any appearance that either is better can only be an unwarranted value illusion.

---

1 Tim Maudlin (2008) discusses the case for allowing the boundaries between grades to vary from context to context, when grading students’ work.

2 More concretely, say the quality of evidence is a matter of its credence, or subjective probability, and let the collective credence in a body of evidence be the average credence of the two graders. There are circumstances under which this makes good sense. Now, suppose the first grader counts evidence as high quality if his credence in it falls within the interval [0.5, 1]. The second grader, applying a higher standard, interprets high, medium and low as the intervals [0.9,1], [0.7,0.9) and [0.5,0.7). Then the situation might be this:
Sometimes, when it is known in advance exactly what it is that grades are supposed to measure, we can hope to avoid all uncertainty about what they mean to different people. If it is opposition to gun-control legislation, for instance, we can specify exactly which proportion of nay-votes will correspond to which grade. We cannot always pin down the meaning of grades, though. When evaluating research proposals, for example, there are different criteria to consider. They include the significance of the proposed research, the reputations of the investigators, and the technical quality of the proposal itself. With restaurants, relevant criteria are the quality of ingredients, mastery of technique, value for money and so on. Now, different ways of combining evaluations by these several criteria will determine different notions of overall merit and, without fixing weights or priorities in advance, it cannot be clear which of these the grades are supposed to measure—let alone precisely which intervals of a single dimension of overall merit the different grades are supposed to pick out. In such cases there can be no precise interpretation of grades.

When for whatever reason a precise interpretation cannot be had, or it is undesirable to have one, it is common to explicate grades using other imprecise expressions. So, for instance, the Arts and Humanities Research Council explains that a Grade 4 research proposal is:

- A very good proposal demonstrating high international standards of scholarship, originality, quality and significance.

And here is what it is for a restaurant to have three Michelin stars:

- Exceptional cuisine where diners eat extremely well, often superbly. Distinctive dishes are precisely executed, using superlative ingredients.

Publishing such descriptions surely can help to promote a common understanding of grades, the more so when people share not only a language but also their culture. Notice though that, by itself, it cannot be expected to remove all uncertainty about how grades are interpreted. One reason for this is that evaluative expressions such as ‘very good’ and ‘extremely well’ are themselves highly contextual. Depending whom you ask, and where and when you ask, ‘distinctive dishes’ and ‘precisely executed’—perhaps even ‘diners eat extremely well’—might cover anything from dinner with Paul Bocuse all the way down to driving through a McDonald’s. Grading restaurants by inviting just

<table>
<thead>
<tr>
<th>Probability/Grade</th>
<th>$E_1$</th>
<th>$E_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>First grader</td>
<td>0.5/high</td>
<td>0.9/high</td>
</tr>
<tr>
<td>Second grader</td>
<td>0.9/high</td>
<td>0.6/low</td>
</tr>
<tr>
<td>Collective</td>
<td>0.7</td>
<td>&lt; 0.75</td>
</tr>
</tbody>
</table>

3 There can be a value illusion in the intended sense even if the appearance is correct. What makes it an illusion is the lack of warrant. Similarly, someone can hallucinate that there is a seal in the bedroom even when there really is a seal in the bedroom.
anyone to go online and click on grade descriptions would presumably not be a rewarding exercise. In practice, expert panels can use several methods to avoid possible equivocation. As well as descriptions there can be training with paradigm examples and protocols for assigning grades.

This is a study of the consequences for collective grading of uncertainty about the interpretation of grades. A first task will be to develop a technical framework in which to give a precise sense to assumptions about the extent of such uncertainty. After that, I shall present some early findings. There is a spectrum of cases to consider. At one end, it has been fixed just how good or bad things have to be for graders to apply the various grades to them, and this is the same for everyone involved. At the other end of the spectrum, while assuming that the graders’ interpretations are coherent, we know nothing substantive about where the lines of application between different grades. I shall report here only on these extremes. In the first case, we will see, collective grading can be, in a technical sense, sound. Intuitively, this means that judgments reached on the basis of grades are warranted by what these grades tell us about merit. In the second case, under certain apparently quite agreeable assumptions, collective grading is inherently unsound. Its judgments are unwarranted value illusions.

I proceed as follows. Section 2 adapts the Arrow-Sen framework for welfare economics to the study of collective grading. There will be, first, a domain of grade profiles. These are hypothetical records of which grades individual people assign to what. Then there will be a separate domain of value profiles. The focus here will be on certain relationships between these two domains. Section 3 develops first the notion of import. The import of a grade profile is a set of value profiles. It models the information about merit that this profile conveys. Section 3 then introduces the notion of soundness and gives an example of a sound procedure for collective grading. Section 4 shows that when their import is too uncertain, grade profiles convey merely ordinal information about merit. Then, as it happens, under certain conditions no collective grading procedures whatsoever are sound. This is a consequence of Kenneth Arrow’s (1951) “impossibility” theorem. Finally, Section 5 considers in light of this result a proposal of Balinski and Laraki (2010) to use collective grading instead of traditional voting methods in political elections.

2. A Framework

Say there are given some items $X$, and also some people $N$ who will assign to them grades from a given set $G$. The goal of collective grading is to arrive at a collective evaluation of the items. We can distinguish two roles in such an exercise. First, there are the graders. Their job is to assign grades to the items, perhaps after discussion among themselves or with others. Second, there is the aggregator, who on this basis of the assigned grades will determine the collective evaluation. The same people can take on both roles. They do, for instance, when two teachers together grade some students’ work. In
other cases the aggregator is a different person, say the chair of an expert panel. With automated online course evaluation, the aggregator is not a person but rather an algorithm.

### 2.1 Grades

A **grading language** is a pair \( (\mathcal{G}, \succ) \), where \( \mathcal{G} \) is a finite set of grades and \( \succ \) a linear ordering of \( \mathcal{G} \). Intuitively, \( \succ \) says which grades are better than which. It is useful to extend the notation: \( g \succ h \) means that either \( g \succ h \) or \( g = h \). Because \( \mathcal{G} \) is finite, \( \succ \) is discrete. That is, there are next best and next worst grades, with nothing in between.

**Example 1:** Let \( X \) be \{a, b, c\} and let \( N \) be \{1, 2\}. And let the grading language be \( \mathcal{G} = \{A, B, C, D, E\} \) with the obvious ordering.

More concretely, \( X \) might be some research proposals, \( N \) two “introducers,” and \( \mathcal{G} \) the five grades that they may assign. Or, in evaluating bodies of evidence, \( A, C \) and \( E \) are primary grades high, moderate and low, with intermediate grades \( B \) and \( D \) there to “split the difference” in case of persisting disagreement among the graders about which primary grade applies.

To study collective grading we can use the Arrow-Sen framework from welfare economics (Arrow 1951, 1963; Sen 1970, 1977). First, where \( i \) is one of the people in \( N \), an **individual grade assignment** is a function \( G_i \) mapping each item in \( X \) to some grade in \( \mathcal{G} \). A **grade profile** \( \langle G_i, \ldots, G_n \rangle \) is a list of these, one for each grader in \( N \), for instance:

**Example 1 (continued):** \( G_1a = A, G_1b = C, G_1c = C; G_2a = C, G_2b = A, G_2c = E \).

A grade profile is a record of how each of the graders in \( N \) might evaluate, by assigning grades in \( \mathcal{G} \), each of the items in \( X \). Grade profiles carry information about how good or bad the graders take things to be. They are the vehicles by which such information passes from the graders to the aggregator. A **grade domain** \( \mathcal{D} \) is a set of grade profiles,\(^4\) representing all such grade reports that the graders might produce, and that the aggregator might be called on to handle. Depending on the case, the domain might be restricted. That is, it might not include all “logically possible” grade profiles. For example, if it so happens that all of the items under consideration are extremely good, the domain might only include profiles in which all of them have high grades. The following example concerns another domain in which the graders use only some of the available grades:

**Example 2:** Let \( X, N, \) and \( \mathcal{G} \) be as in **Example 1**. Let \( \mathcal{D} \) be the set of all grade profiles \( G \) such that for all \( i \in N \), and for all \( x \in X \): \( G_ix \in \{A, C, E\} \).

\(^4\) \( \mathcal{G} \) is for ‘grade’. The superscript distinguishes a grade domain from the **value domains** that I shall soon introduce.
This domain has a property that will be important later on. For any given individual assignment $G_i$, let the binary relation $\tilde{G}_i$, or flat $G_i$, be defined as follows: $x \tilde{G}_i y$ iff $G_i x \succeq G_i y$. Notice that $\tilde{G}_i$ is a weak ordering of $X$, both transitive and complete. It inherits these properties from $\varphi >$. There are ties whenever two items receive the same grade. This idea extends in the obvious way to grade profiles: $\tilde{G} = \langle \tilde{G}_i \rangle$. Now, a grade domain is ordinally unrestricted if for any list $\langle R_i \rangle$ of weak orderings of $X$, there is some $G$ such that $\tilde{G} = \langle R_i \rangle$. The domain $\mathcal{D}_\varphi$ of EXAMPLE 2, though restricted, is ordinally unrestricted.

Let a grading function be a function mapping each grade profile $G$ in its domain onto a partial grade assignment: one that assigns a grade to some items in $X$ but perhaps not to all of them. For any such partial assignment $G_N$, let $x \smile y$ iff $G_N x \not\succeq G_N y$. This relation $\smile$ is a preorder of $X$, a binary relation that is reflexive and transitive but not necessarily complete. Below, functions mapping grade profiles onto preorders will also be called grading functions. It will be useful to use $\lambda$-notation to refer to them. Thus $\lambda G. \smile$ will stand for a function that maps each $G$ in its domain onto a corresponding $\smile$.

Grading functions are a rendering in this framework of the role of aggregator.

Incompleteness in $\smile$ might be found unattractive when the object of evaluating things is to support a choice among them. If there are three research proposals under consideration, for instance, and two can be funded, an outcome is to be avoided in which we count one as better than both of the others, but cannot compare the remaining two with respect to one another. The reason for allowing aggregation to introduce indeterminacy is that the set of grades is finite. When one grader thinks an item merits an $A$, and the other only a $B$, and there is no available grade in between, there might be no saying what the outcome should be.

Sometimes it will be possible to achieve completeness by making sure that there are enough grades available to “split the difference” when graders disagree. The following describes a fortunate case in which there are already enough grades.

EXAMPLE 3: Let $\mathcal{D}_\varphi$ be as in EXAMPLE 2. Define a grading function $\lambda G. \preceq$ as follows. First, for any $g \in G$, let $middle\{g, g\}$ be $g$. Second, for any distinct grades $g_i$ and $g_j \in G$, suppose there is some further grade $g_2$ that lies, in the sense of $\succ$, in the middle between $g_i$ and $g_j$. Then $middle\{g_i, g_j\}$ is this $g_2$. If on the other hand there is no such $g_2$ then $middle\{g_i, g_j\}$ is undefined. With $G = \{A, B, C, D, E\}$, for instance, $middle\{D, D\} = D$, $middle\{A, C\} = B$ and $middle\{E, A\} = C$; but $middle\{B, C\}$ and $middle\{B, E\}$ are undefined.

Now, for any given profile $G$, let $x \preceq y$ hold iff $middle\{G_i x, G_i y\}$ and $middle\{G_j x, G_j y\}$ are both defined, and $middle\{G_i x, G_j x\} \preceq middle\{G_i y, G_j y\}$.

---

5 Preorders are also called quasiorderings.
Then $\lambda G \geq$ is a grading function for $\mathcal{D}$. To be shown is that for any $G \in \mathcal{D}$, $\geq$ is a reflexive and transitive relation on $X$. Note that because $G_x \in \{A, C, E\}$, for each $i$ and $x$, middle$\{G_i x, G_x\}$ is always defined. Reflexivity and transitivity of $\geq$ follow immediately from the corresponding properties of $\succ$. Similarly, the completeness of $\succ$ ensures that $\lambda G \geq$ satisfies:

**Completeness (C):** for each $G \in \mathcal{D}$, $\geq$ is a complete relation.

Here are some further conditions. $\lambda G \geq$, with domain $\mathcal{D}$, satisfies:

*Ordinally Unrestricted Domain (OU)* iff $\mathcal{D}$ is ordinally unrestricted;

*Weak Pareto (WP)* iff for any $x, y \in X$ and $G \in \mathcal{D}$: if for each $i \in N$, $G_i x \succ G_i y$, then $x \succ y$;

*Non-Dictatorship (D)* iff there is no $\delta \in N$ such that for any $x, y \in X$ and $G \in \mathcal{D}$: if $G_\delta x \succ G_\delta y$ then $x \succ y$;

*Independence of Irrelevant Alternatives (I)* iff for any $x, y \in X$ and profiles $G, H \in \mathcal{D}$: if $G \{x, y\} = H \{x, y\}$, then $G \geq \{x, y\} = H \geq \{x, y\}$.\(^6\)

The discussion of **THEOREM 13** in Section 4 refers to the fact that the grading function of **EXAMPLE 3** has some of these properties. It is easily seen to have them all:

**EXAMPLE 3 (CONTINUATION):** $\lambda G \geq$ satisfies:

*OU*: since there are just three items in $X$, any weak ordering of $X$ is induced by some assignment to these items of the three grades $A, C$ and $E$.

*WP*: consider any $x, y \in X$ and suppose, for both $i$, $G_i x \succ G_i y$. Since $G_i$ only assigns grades in $\{A, C, E\}$, there are just 9 ways this can happen. In each case it is easily checked that $x \succ y$;

*D*: Consider any profile $G$ such that $G_1 a = G_2 b = A$ and $G_1 b = G_2 a = C$. We have $G_1 a \succ G_1 b$, and $G_2 b \succ G_2 a$. Since middle$\{G_1 a, G_2 a\} = middle\{G_1 b, G_2 b\} = B$, though, we have neither $a \succ b$ nor $b \succ a$. So neither $1$ nor $2$ is a dictator of $\lambda G \geq$;

---

\(^6\) The restriction of a function to some subset of its domain, and more generally that of a list of functions to some subset of their common domain, has the obvious meaning. So $G \{x, y\} = H \{x, y\}$ means that for all $i$, $G_i x = H_i x$ and $G_i y = H_i y$. 

7
I: Obviously, from the definition of $\lambda G \succeq$, the comparison of two items by $\succeq$ depends only on the grades assigned to these two by $G$.

The observations of Example 3 (Continuation) are noteworthy because these are versions of the conditions of Arrow’s (1951) theorem, adapted to the richer informational setting in which people do not merely rank alternatives but instead grade them. Since $\lambda G \succeq$ satisfies all of these conditions, it might appear that Arrow’s theorem does not threaten collective grading. One main point of the coming Sections is that it still does indirectly, when there is too much uncertainty about what grades mean to individual graders, and the extent to which this is coordinated from one grader to the next. For, as we will see in Section 4, Arrow’s theorem in this case entails that, subject to the conditions in question, collective grading cannot be sound, in the sense to be made precise in Section 3. Before we get to any of this, though, we must first turn to the matter of what grades mean.

2.2 Values

It is time to bring in the dimensions of goodness, quality and merit that grades measure. The term ‘value’ will cover all these. A value structure is a pair $(\mathcal{V}, \triangleright)$, where $\mathcal{V}$ is a set and $\triangleright$ is a linear ordering of $\mathcal{V}$ (from better to worse). We will assume that $\triangleright$ is dense, so that for any two values there is another in between. For example, in connection with grading research proposals, $\mathcal{V}$ is a dimension of “overall” merit gotten by somehow weighing or prioritizing the different criteria of quality. When grading evidence, $\mathcal{V}$ might be a real interval representing subjective probabilities.

I shall now again borrow the Arrow-Sen framework by introducing, in addition to the grade domain, the value domain, the value domain $D^\mathcal{V}$ is a set of value profiles. The role of these notions, though, is somewhat different here. Value profiles are not supposed to represent any actual or possible report on the merit of things. That is what grade profiles are for. Rather, they will contribute to an account of the content

---

7 For a statement of these conditions and of Arrow’s theorem itself, see List 2013.
of such reports—of the information that grade profiles convey from graders to aggregator. Value profiles represent hypothetical estimations of the value of things.

We will assume that value domains are unrestricted, which means that they include every list of individual value assignments for the relevant $X$, $N$, and $V$. Because value domains play a purely theoretical role, this assumption does not have any obvious empirical consequences for the set of items under consideration, or how the graders determine which grades to assign to them, or anything like that. Since all value domains are unrestricted, it will often not be necessary to say explicitly which one is meant. By fixing $X$, $N$, and $V$ we will have fixed that as well.

A value function maps each value profile $V$ in its domain onto a (complete) value assignment. It determines in the obvious way a function $\lambda V. v \geq$ mapping value profiles onto weak orderings $v \geq$ of $X$, and these too will be called value functions.

**Example 4:** Let $X$ and $N$ be as in the previous examples. Let $V$ be the real interval $(0,1)$ with the obvious ordering. The value domain $\mathcal{D}^V$ is then $\{<V_i, V_j>: V_i: X \rightarrow (0,1)\}$. The following specifies a value function $\lambda V. v \geq$ for $\mathcal{D}^V$. For any $<V_i, V_j> \in \mathcal{D}^V$, and for any $x,y \in X$, we put:

$$x V_i \geq y \iff V_i x + V_j y / 2 \geq V_i y + V_j x / 2.$$ 

Here, $v \geq w$ if either $v > w$ or $v = w$. Later, in Example 7 (continuation) we will have another but closely related value function for this domain.

## 3. Sound grading procedures

This Section introduces the notion of a sound procedure for collective grading. A grade profile, recall, is a report of which items have received which grades from whom. In outline, a grading function is sound if the judgments made using it are, given the information about merit that a grade profile conveys, warranted. Section 3.1 takes up the question of which information this is. Section 3.2 defines the notion of soundness and gives an example of a sound grading function.

### 3.1 Import

We will need an account of the information about merit that a grade profile $G$ carries from the graders to the aggregator. To this end I shall introduce the notion of import. The import of $G$, written $[G]$, will be a set of value profiles. Intuitively, the import of a grade profile includes all hypothetical estimations of merit that are compatible with the reported grades. The less the aggregator knows about the

---

8 Knows, believes, assumes, supposes…. Here it doesn’t matter which.
graders’ understanding of the grades, the less information $G$ conveys, and the more possibilities there are in $\llbracket G \rrbracket$.

The basic notion here is that of an interpretation of a grading language. An interpretation tells us what any given grade signifies, by assigning to it an interval of some dimension of value. Technically:

**Definition 5:** An interpretation $I$ of $(G, \varphi^>)$ in $(V, \triangleright)$ is a function $I : G \rightarrow \mathcal{P}ow(V)$ such that:

For each $g \in G$, $I(g)$ is convex: if $v_1, v_2 \in I(g)$ and $v_1 \triangleright v_2 \triangleright v_3$, then $v_2 \in I(g)$; and:

$I$ is orderly: for any $g_1, g_2 \in G$, $v_1 \in I(g_1)$, and $v_2 \in I(g_2)$, if $g_1 \varphi^> g_2$ then $v_1 \triangleright v_2$; and:

$\{I(g) : g \in G\}$ is a partition of $\triangledown$ for any $g_1, g_2 \in G$, $I(g_1) \cap I(g_2) = \{\}$, and $\cup_{g \in G} I(g) = \triangledown$.

There is no need to assume that it is logically possible to earn any given grade. To make clear that nothing turns on this, though, it is better to insist that it is so:

For each $g \in G$, $I(g) \neq \{\}$.

All teachers are familiar with interpretations from the keys or scales used to turn percentage scores on tests into corresponding letter grades. Here is an example:

**Example 6:** Let $G = \{A, B, C, D, E\}$ and $V = (0,1)$. $I$ interprets $G$ within $V$ as follows: $I(A) = [0.8,1)$, $I(B) = [0.6,0.8)$, $I(C) = [0.4,0.6)$, $I(D) = [0.2,0.4)$ and $I(E) = (0,0.2)$.

Let $I$ interpret $(G, \varphi^>)$ within $(V, \triangleright)$, and let $G_i : X \rightarrow G$ be the grade assignment of some individual $i$. A next notion is the import $\llbracket G_i \rrbracket^I$ of $G_i$, relative to $I$. Intuitively, this is the set of all hypothetical assignments of value that are compatible with $G_i$, assuming that all grades are interpreted according to $I$. That is, $\llbracket G_i \rrbracket^I$ is the set $\{V' : X \rightarrow \triangledown^I$ for each $x \in X, Vx \in I(Gx)\}$.

The import $\llbracket G_i \rrbracket^I$ of $G_i$ is a measure of what $G_i$ conveys to the aggregator, if the aggregator assumes that $i$’s understanding of the grades is as specified by $I$. The import $\llbracket G \rrbracket$ of an entire grade profile $G$ will now be seen as value profiles, and will measure the information conveyed by $G$.

**Example 7:** Let $X = \{a, b, c\}$, let $N = \{1, 2\}$, let $G = \{A, B, C, D, E\}$, let $V = (0,1)$, and let $I$ be an interpretation of $G$ in $V$. Consider any suitable grade domain $\mathcal{D}^V$, and let $\mathcal{D}^V$ be the (unrestricted) value domain for $X, N$ and $\triangledown$. For any given $G \in \mathcal{D}^V$, let $\llbracket G \rrbracket$, the import of $G$ within $\mathcal{D}^V$, be:

$$\{<V_1, V_2> \in \mathcal{D}^V : V_1 \in \llbracket G_1 \rrbracket^I \text{ and } V_2 \in \llbracket G_2 \rrbracket^I\}.$$  

The import function $\lambda G.\llbracket G \rrbracket$ of Example 7 embodies a strong assumption about the grading exercise at hand: the aggregator knows that each grader has settled on an interpretation of the grades, the same one for both of them, and knows also which interpretation this is: $I$. Now I shall distinguish several
kinds of import function. They embody different assumptions about the aggregator’s knowledge. Let \( \mathcal{S} \) be the set of all interpretations of a given grading language in some value structure.

**Definition 8:** An import function \( \lambda G, [G] \) is:

- **Grounded and rafted** iff there is some \( I \in \mathcal{S} \) such that for every \( G \in \mathcal{S}_I \), \([G] = \times_{i \in N} [G]_i^I \);

- **Grounded and unrafted** iff for each \( i \in N \) there is some \( I_i \in \mathcal{S} \) such that for every \( G \in \mathcal{S}_{I_i} \), \([G] = \times_{i \in N} [G]_i^I \);

- **Floating and rafted** iff for every \( G \in \mathcal{S}_I \), \([G] = \bigcup I \in \mathcal{S} \times_{i \in N} [G]_i^I \);

- **Floating and unrafted** iff for every \( G \in \mathcal{S}_I \), \([G] = \times_{i \in N} \bigcup I \in \mathcal{S} [G]_i^I \).

The import of a grade profile is said to be grounded or floating, rafted or unrafted, accordingly as the import function is. To illustrate, the import of the grade profiles in Example 7 is grounded and rafted. When the import of \( G \) is floating and unrafted, \( V \in [G] \) if there are any interpretations \( I_1, \ldots, I_n \) such that, for any individual \( i \), \( V_i \in [G]_i^I \). Intuitively, what this means is that, as far as the aggregator can tell, the graders might interpret the grades any coherent way at all, and they might do so quite independently of one another.

The distinctions of Definition 8 cover only a few kinds of uncertainty about the applicability of grades. For one thing, they presuppose a fixed dimension \( \mathcal{V} \) of value. All interpretations in \( \mathcal{S} \) concern this same one. As I mentioned in Section 1, though, when there are several evaluation criteria to consider, and the weights or priorities of the different ones have not been settled in advance, there will not be any single dimension of overall quality that is uniquely the one being measured. In these cases, grades will be not only coarse grained but also vague. That is, not only will any given grade apply to items of different merit; in addition, there will be **borderline cases** in which it is indeterminate whether this grade applies. There is no need to go further into the vagueness of grades, here, important and interesting though this matter is. Any additional uncertainty that it introduces can only compound the problem of value illusions.

### 3.2 Soundness

Finally, we have the notion of a sound grading procedure:

**Definition 9:** \( \lambda G, G \geq \) is sound with respect to \( \lambda V, V \geq \), assuming \( \lambda G, [G] \), iff:

For all \( G \in \mathcal{S}_I \) and all \( x, y \in \mathcal{X} \): (a) if \( x \geq y \) then for every \( V \in [G] \), \( x \geq V \); and (b) if \( x > y \) then for every \( V \in [G] \), \( x > V \).
Questions about the soundness of grading procedures will often arise in connection with some given assumptions about how people understand their language, and in particular grades. These assumptions are embodied, we may suppose, in a contextually fixed import function, \( \lambda G \texttt{[} G \texttt{]} \). Then we can enquire simply into the soundness of any given \( \lambda G \texttt{[} G \texttt{]} \) with respect to any given \( \lambda V \texttt{[} V \texttt{]} \). The grading function \( \lambda G \texttt{[} G \texttt{]} \) is simply said to be *sound* if there is some \( \lambda V \texttt{[} V \texttt{]} \) with respect to which it is sound,\(^9\) and otherwise it is *unsound*.

To see what soundness amounts to, consider an example. We are going to rank some research proposals on the basis of grades assigned to them by the members of an expert panel. We have provided everyone with descriptions of the applicable grades and we know there is some correspondence between the graders’ estimation of the merit of the proposals and the grades they have assigned to them. Now we receive the report from the panel, \( G \). Grades, of course, are coarse grained measurements of merit. And there remains some uncertainty about how the different members of the panel have applied them. So this report is compatible with a range of different lists \( V \) of graders’ estimations of the merit of the proposals. An import function will tell us which ones these are: \( \texttt{[} G \texttt{]} \). Fix an import function \( \lambda G \texttt{[} G \texttt{]} \) embodying what we know, assume or presuppose about how the graders interpret the grades.

Now, suppose our collective grading procedure \( \lambda G \texttt{[} G \texttt{]} \) is sound with respect to some particular function \( \lambda V \texttt{[} V \texttt{]} \). And suppose it reaches a verdict on two proposals: they are equally good, say. Then this comparison of this pair by \( \texttt{[} G \texttt{]} \) is guaranteed to agree with the comparison \( \texttt{[} V \texttt{]} \) obtained using \( \lambda V \texttt{[} V \texttt{]} \), for any \( V \) at all that is compatible with \( G \). That is, the judgment that \( \lambda G \texttt{[} G \texttt{]} \) hands down on the basis of the assigned grades alone will be confirmed when we calculate, using \( \lambda V \texttt{[} V \texttt{]} \), the aggregate estimated merit of the proposals.

Soundness is desirable because decisions made using sound procedures can be rationalized. If someone wants to know why their proposal wasn’t funded then, if the grading procedure was sound, there is an answer. We can say that *this* \( G \) was the grade report from the expert panel; that *this* \( \texttt{[} G \texttt{]} \) is what the report tells us about how the graders’ might have estimated the different proposals’ merit; that *this* ranking \( \texttt{[} V \texttt{]} \) of the proposals, obtained by aggregating any one \( V \) of the possible estimations, agrees with the ranking \( \texttt{[} G \texttt{]} \) that we reached just on the basis of the grades—and that their proposal sadly just wasn’t close enough to the top.\(^{10}\)

---

\(^9\) Similarly, someone is married if there is some person to whom they are married.

\(^{10}\) To rationalize the decision is not the same thing as justifying it, because different parts of the rationalization are themselves subject to reasonable criticism. The disgruntled can argue that our assumptions about what the assigned grades \( G \) say about merit, embodied in \( \lambda G \texttt{[} G \texttt{]} \), are incorrect (“that is *not* how the panel members understood the grades!”); or that agreement with \( \texttt{[} V \texttt{]} \) is beside the point, because the procedure \( \lambda V \texttt{[} V \texttt{]} \) for aggregating estimations of merit that we used is unsuitable (“some of the members are not experts, and their views do not carry the same weight!”). Or they can argue that both \( \lambda G \texttt{[} G \texttt{]} \) and \( \lambda V \texttt{[} V \texttt{]} \) are inappropriate because,
EXAMPLE 7 (CONTINUATION): More specifically, let $\mathcal{D}$ be as in EXAMPLE 2, and let $I$ be as in EXAMPLE 6. Consider again the grading function $\lambda G, \geq$ for $\mathcal{D}$ from EXAMPLE 3. Now let the value function $\lambda V, \geq$ be defined as follows. For any $V \in \mathcal{D}$ and $x, y \in X$, we put:

$$x^V \geq y \iff I^1(V_1x + V_2x / 2) \geq I^1(V_1y + V_2y / 2)$$

The inverse $I^{-1}$ of $I$ is obvious but not standard: e.g. $I^{-1}(0.5) = C$. Now we will see that $\lambda G, \geq$ sound with respect to $\lambda V, \geq$, assuming $\lambda G, [G]$ (the import function defined earlier in EXAMPLE 7). This follows directly from:

FACT: for any $G \in \mathcal{D}$, $x, y \in X$, and $V \in [G]$: $x^G \geq y$ if and only if $x^V \geq y$.

To verify FACT, note first that for any $g_1, g_2 \in \{A, C, E\}$, and for any $v_1 \in \mathcal{I}(g_1)$ and $v_2 \in \mathcal{I}(g_2)$:

$$I^1(v_1 + v_2 / 2) = middle\{g_1, g_2\}.$$

To verify this just check all the possibilities; really, there are only three. Now consider any $V \in [G]$. By choice of the import function $\lambda G, [G]$, for any $x, y \in X$ and $i \in N$ we have $V_1x \in \mathcal{I}(G_iY)$ and $V_2y \in \mathcal{I}(G_iY)$, so in particular:

$$I^1(V_1x + V_2x / 2) = middle\{G_1X, G_2X\},$$

and:

$$I^1(V_1y + V_2y / 2) = middle\{G_1Y, G_2Y\}.$$ 

Therefore we have:

$$x^G \geq y \iff middle\{G_1x, G_2x\} \nleftarrow\rightarrow middle\{G_1y, G_2y\} \iff I^1(V_1x + V_2x / 2) \nleftarrow\rightarrow I^1(V_1y + V_2y / 2) \iff x^V \geq y.$$

This completes the demonstration of the FACT, and with it the soundness of $\lambda G, \geq$ with respect to $\lambda V, \geq$, assuming $\lambda G, [G]$.

It is relevant to the discussion of THEOREM 13 in the next Section that the value function $\lambda V, \geq$ satisfies Independence of Irrelevant Alternatives. That is, for any $x, y \in X$, and any $V, W \in \mathcal{D}$: if having given the wrong weights or priorities to the different criteria for evaluating proposals, we are working with the wrong dimension $\forall \nu$ of overall merit, and therefore within altogether the wrong value domain. Still, the rationalization is a beginning. It provides the main structure for a justification of our grading exercise, and indicates the points that might need support.
\[ V(x',y') = W(x',y'), \text{ then } V \geq W \geq V. \] That it does is obvious because for any \( V \in \mathcal{D} \), whether \( x' \geq y \) depends only on the values assigned to \( x \) and \( y \).

The import function of EXAMPLE 7 is grounded and rafted. It embodies strong assumptions about the graders’ understanding of the grades. In addition, as we have seen in EXAMPLE 7 (CONTINUATION), the grading function \( \lambda G.\geq \) is sound, on these assumptions. Now we will see that this matter of import is crucial to the existence of sound grading functions.

4. When collective grading can only create illusions

Uncertainty about what grades mean to different people seriously undermines collective grading. It does so by destroying the information about merit that grades convey. In extreme cases, as we will see, there is only ordinal information left. People might then as well forget the finer distinctions of grading and just rank things from better to worse. Then all grading functions are, under certain minimal conditions, unsound.

Recall that \( \bar{G}_i \) is the weak ordering of \( X \) determined by an individual grade assignment \( G_i \). \( \bar{G}_i|S \) stands for the weak ordering of any given \( S \subseteq X \). The following lemma plays a central part in the coming discussion of the destruction of information:

**Lemma 10:** Let \( S \subseteq X \), and let \( G_i \) and \( H_i \) be two individual grade assignments such that \( \bar{G}_i|S = \bar{H}_i|S \).

Letting \( I \) be any interpretation of the grading language, choose any \( W_i \in \llbracket H_i \rrbracket^I \). Then there is some interpretation \( I \), and some \( V_i \in \llbracket G_i \rrbracket^I \), such that \( V_i|S = W_i|S \).

It is here that we use the assumptions that \( \triangleright \) is dense, and that there is no minimal or maximal value in \( \mathcal{V} \). For a proof sketch, see the Appendix. Putting \( S = X \), LEMMA 10 has the direct consequence:

**Corollary 11:** Suppose \( \bar{G}_i = \bar{H}_i \). Then \( \bigcup_{i \in I} \llbracket G_i \rrbracket^I = \bigcup_{i \in I} \llbracket H_i \rrbracket^I \).

Now we will see the destructive force of uncertainty about the meaning of grades. From DEFINITION 8 and COROLLARY 11 we have:

**Corollary 12:** Suppose the import function \( \lambda G.\llbracket G \rrbracket \) is floating and unrafted. If \( \bar{G} = \bar{H} \), then \( \llbracket G \rrbracket = \llbracket H \rrbracket \).

This tells us that any two grade profiles that determine the same orderings have the same import. That is, when there is this much uncertainty about how grades are being interpreted, grade profiles convey

---

\[ 11 \] I haven’t written out all details of the proof yet but feel confident that it is correct.
only ordinal information about merit. Aggregation of ordinal information about preferences runs up against a firm theoretical barrier in Arrow’s (1951) theorem. Here is a closely related limit to the possibilities for meaningful collective grading:

**THEOREM 13:** Let \( X \) include at least three items, and let \( \lambda_G[G] \) be floating and unrafted. Let \( \lambda_G[G] \geq \) be a grading function that satisfies *Ordinally Unrestricted Domain, Completeness, Weak Pareto* and *Non-Dictatorship*, and let \( \lambda_V[V] \geq \) be a value function that satisfies *Independence of Irrelevant Alternatives*. Then \( \lambda_G[G] \geq \) is not sound with respect to \( \lambda_V[V] \geq \) assuming, \( \lambda_G[G] \).

For a proof, see the Appendix. This result places a severe limit on the possibilities for meaningful collective grading when too little is known about how individual graders understand grades, and the extent to which their understanding is coordinated from one person to the next. Under the given conditions, procedures for collective grading *necessarily* are unsound.

**EXAMPLES 3 AND 7** showed that the grading and value functions of **EXAMPLE 7** satisfy all of the assumptions of **THEOREM 13**, apart from the one about the import of grade profiles. Evidently, the possibilities for sound grading really do turn on assumptions about import.

The next Section considers in light of **THEOREM 13** the prospects for collective grading by certain large and diverse groups of people: the electorates of modern democracies.

5. Grading and Voting

…the great Augustus himself, in possession of that power which ruled the world, acknowledged he could not… appoint what idea any sound should be a sign of, in the mouths and common language of his subjects. John Locke, in *An Essay Concerning Human Understanding*\(^{12}\)

Voters rank the candidates in some election from better to worse on a ballot, and then their rankings are compiled by some method into a collective ranking. The candidate on top of this collective ranking wins. Kenneth Arrow’s “impossibility” theorem describes a theoretical limitation: no voting method can satisfy a short list of desiderata (Arrow 1951). Some have taken his result to show that there can in fact be no such thing as democracy, conceived as government by a common will of the people, as revealed by voting. Supposedly the very idea of that, captured by the desiderata, is incoherent.\(^{13}\)

Others say that Arrow’s theorem is about the impossibility of trying to do too much with too little information. On this analysis, the lesson to be drawn is not that democracy is impossible but rather that more information is needed to arrive at a collective ranking than just voters’ rankings of the candidates. Now, by grading you can convey more than just a ranking. Having said that one candidate

\(^{12}\) Locke 1961 (1690), Vol. 2, pp. 14-15

\(^{13}\) See Riker 1985.
is “excellent,” but another merely “good,” you convey that one is better than the other; you can convey more than just this, though, because different grades determine the same ranking: had you said that the second candidate is “unacceptable,” the implied ranking would have been the same. Intriguingly, the information in grades can be enough to dissolve some of the problems with traditional voting methods, including Arrow’s theorem. That emerges here from Example 3 (continuation). In a recent book, Michel Balinski and Rida Laraki (2010) argue that we should stop using these methods in elections and start grading candidates instead.

There is a serious difficulty with their proposal.

Collective grading has typically been used in more or less formal settings, by small groups of highly trained people working in close coordination with one another. Look at the list at the beginning: there are expert panels evaluating project proposals. There are professional Michelin inspectors with years of training who get together at “star meetings” to grade restaurants. There are specialists who over decades have developed elaborate protocols for evaluating bodies of evidence in medicine. Balinski and Laraki provide many further examples of this sort. Now, when there is so much shared training and culture there is bound already to be a good common understanding of which dimensions of value are being measured, and of how grades are to be understood. Also, there will be plenty of opportunities to cultivate further such coordination if need be. In addition to linguistic definitions, there can be examples of correctly assigned grades, elaborate protocols to be followed when assigning grades, and more. In Section 3, we found reason to expect that, when the application of grades is this tightly constrained, collective grading can be meaningful.

Voting, on the other hand, has often used with much looser collections of people. Certainly in large, culturally and linguistically diverse democracies such as those of India or the United States we should expect not only that there will be a great need for instruction in the application of grades, if collective grading is not to devolve into equivocation, but also rather few opportunities to cultivate a common understanding. Theorem 13 tells us that, when the application of grades is insufficiently constrained, collective grading can only create value illusions. In light of this it seems likely that, unless something can be done to settle what grades signify “in the mouths and common language” of the people, collective grading will not be an acceptable procedure for political elections in modern democracies.

Now, Balinski and Laraki are aware that if collective grading is to be used in elections then the people will need a common understanding of the applicable grades. They argue that these should be formulated in a language that is shared by the voters. Since the evaluative expressions of natural languages are highly contextual, though, it seems unlikely that this will be enough. Instead of bringing the problem of uncertain import any further forward to a solution, it would appear that their proposal

---

14 To gain an impression of how such agreement is cultivated when grading medical evidence, see (Balshem et al. 2011).
merely shunts it off sideways, to the contextuality of this language and consequent uncertainty about how voters will understand grade descriptions couched in its terms.

I conclude that it remains to be seen whether collective grading can be a meaningful method for political elections in large democracies. Linguistic and psychological studies into language learning, categorization and elicitation methods might give us some insight into how an electorate’s understanding of evaluative expressions could be coordinated. Realistically, we might hope to succeed in this to some extent but not completely. Further analytical work into the consequences of uncertain import, in cases other than the two extremes considered here, might give us some idea of how much coordination will be needed.
Appendix

Proof sketch for Lemma 10: Assuming \( G_i|S = H_i|S, G_i \) and \( H_i \) determine the same weak ordering of \( S \). It can be represented as a finite series of cells \( C_{gh} \), each labeled by the grades that \( G_i \) and \( H_i \) assign to the elements within. That is, \( C_{gh} = \{ s \in S : G_s = g \} = \{ s \in S : H_s = h \} \). Now take any interpretation \( J \) and individual value assignment \( W_i \in [H_i]^I \). The idea of the proof is to read off, from \( W_i \) and the different cells \( C_{gh} \), precursors \( V_i \) and \( I \) of the sought value function \( V_i \) and interpretation \( I \), for which \( V_i \in [G_i]^I \) and \( V_i|S = W_i|S \). They are precursors in that, while not yet themselves a value function and an interpretation, they have properties such that, when suitably expanded into these, they will fill the bill.

The precursor \( V_i \) is obtained by putting, for each \( s \in S \), \( V_i(s) = W_i(s) \). It is not yet a value function because it doesn’t in general assign values to all items in \( X \). Plainly, though, \( V_i \) is already such that \( V_i|S = W_i|S \). The precursor \( I \) is obtained by letting, for any \( g \) assigned by \( G_i \) to some \( s \in S \), \( I_g \) be the least-inclusive interval of \( V \) that includes all of \( \{ W_t : t \in C_{gh} \} \). It is not yet an interpretation because in general it doesn’t assign intervals to all grades, and the intervals it does assign do not together exhaust \( V \). But \( I \) already has some required properties. For instance, its values are convex (because they are intervals) and it can be shown to be orderly (because \( J \) is). Furthermore, we have the beginnings of the required property that \( V_i \in [G_i]^I \), which means that for all \( x \in X \), \( V_i(x) \in IG \). For we have, for all \( s \in S \), \( V_i(s) \in IG \). This can be seen as follows: for any given \( s \in S \) (and letting \( g \) be such that \( G_s = g \)) we have: \( V_i(s) = W_i(s) \in \{ W_t : t \in C_{gh} \} \) (by definition of \( C_{gh} \)) = \( I_g \) (by definition of \( I \) = \( IG \)).

Turning these precursors into \( I \) and \( V_i \in [G_i]^I \) such that \( V_i|S = W_i|S \) requires in the first instance choosing suitable intervals of \( V \) for grades not yet interpreted by \( I \) and choosing suitable values in \( V \) for items not in \( S \), so that these properties of the precursors are maintained. It is in choosing these intervals and values that we will use the assumptions concerning the structure of \( V \), namely that it is densely ordered by \( \triangleright \) and that there are no minimal or maximal elements. These assumptions ensure that there are ‘unused’ intervals to choose. A last step in the construction of \( I \) will be to expand the preliminary intervals of \( V \) that have been assigned to the different grades, so as to end up with a partition of \( V \).

This completes this sketch of a proof of Lemma 10.

Proof of Theorem 13: If \( G \in D \) then its flattening \( \hat{G} \) is a profile, in the sense of Arrow (1951), because for each \( i \) the component \( \hat{G}_i \) is a weak ordering of \( X \). Let \( D^{arrow} = \{ \hat{G} : G \in D \} \). \( D^{arrow} \) is a domain, in Arrow’s sense. The proof will proceed by assuming that \( \lambda G \hat{G} \geq \) is sound with respect to \( \lambda V \hat{V} \geq \) and then constructing from \( \lambda V \hat{V} \geq \) the analogue of a social welfare function \( f \) for \( D^{arrow} \). Under the assumptions of Theorem 13, the constructed \( f \) may then be seen to satisfy the conditions of Arrow’s...
(1951) theorem (see for example List 2013). Since by Arrow’s theorem there can be no such $f$, this
shows the initial soundness assumption to be false.

Suppose then that are at least three items in $X$, and that the import function $\lambda G.\llbracket G \rrbracket$ is floating and
unrafted. Let $\lambda G.\geq$ be a grading function that satisfies Ordinally Unrestricted Domain, Completeness,
Weak Pareto and Non-Dictatorship, and let $\lambda V.\geq$ be a value function that satisfies Independence of
Irrelevant Alternatives. Finally, suppose for the contradiction that $\lambda G.\geq$ is sound with respect to
$\lambda V.\geq$. First we have:

**FACT 14:** For any $G \in \mathcal{D}$ and $V, W \in \llbracket G \rrbracket$, $V \geq W \iff V \geq W$.

Proof: Letting $V, W \in \llbracket G \rrbracket$, suppose for any given $x, y \in X$ that $x \geq y$. We have $x \geq y$, since otherwise by
Completeness of $\lambda G.\geq$, and therefore of $G \geq$, we would have $y \succ x$, and so by part (b) of DEFINITION
9 also $y \succ x$, contradicting the initial supposition that $x \geq y$. By part (a) of DEFINITION 9, since $x \geq y$
also $x \geq y$. Similarly, if $x \geq y$ then $x \geq y$.

Notice that this fact only requires the soundness and completeness properties. Now we will construct $f$, the
analogue of a social welfare function for $\mathcal{D}^{\text{unrow}}$. For each profile $\hat{G} \in \mathcal{D}^{\text{unrow}}$, we first choose some
$V \in \llbracket G \rrbracket$ and put: $f \hat{G} = V \geq$. By FACT 14, it doesn’t matter which $V$ we choose. For $f$ to be a function it
must be that $f \hat{G} = f \hat{H}$ whenever $\hat{G} = \hat{H}$, but that is ensured by COROLLARY 12 which tells us that,
because $\lambda G.\llbracket G \rrbracket$ is floating and unrafted, in this case $\llbracket G \rrbracket = \llbracket H \rrbracket$.

It remains only to verify that $f$ satisfies the conditions of Arrow’s theorem: **Universal Domain (U),
Ordering (O), Weak Pareto (WP), Non-Dictatorship (D), and Independence of Irrelevant Alternatives
(I)** (see List 2013).

**U:** Consider any list $<R>$ of $n (= |N|)$ weak orderings of $X$. Since by Ordinally Unrestricted
Domain the domain $\mathcal{D}$ of $\lambda G.\geq$ is ordinarily unrestricted, there is some $G \in \mathcal{D}$ such that $\hat{G} =<R>$. So $<R> \in \mathcal{D}^{\text{unrow}}$.

**O:** $f \hat{G}$ is a weak ordering of $X$ because, by definition of a value function, $V \geq$ always is.

**WP:** Suppose for each $i$, $x_i \geq y_i$ but not $y_i \geq x_i$. Then, for each $i$, $G_i \not\succ G_{i'}$, and since by assumption
$\lambda G.\geq$ satisfies Weak Pareto we have $x_i \geq y$. Now consider any $V \in \llbracket G \rrbracket$. By soundness of $\lambda G.\geq$
with respect to $\lambda V.\geq$, we have $x \geq y$. So by choice of $f$ and FACT 14, $x f \hat{G} y$ but not $y f \hat{G} x$.

**D:** Suppose for contradiction that $\delta$ is a dictator of $f$. We will see that $\delta$ is a dictator of $\lambda G.\geq$, which
by our assumption of Non-Dictatorship it cannot be. To this end, take any $G \in \mathcal{D}$, and any
$x, y \in X$, and suppose that $G x \not\succ G y$. Then $x \hat{G} y$ but not $y \hat{G} x$, and so, since $\delta$ is a dictator of $f$, $x$
f $\hat{G} y$ but not $y f \hat{G} x$. Choose any $V \in \llbracket G \rrbracket$. By FACT 14 and choice of $f$ we have $x \geq y$. By soundness
therefore $x \geq y$ (since otherwise because $\geq$ is complete we have $y \geq x$ and, by part (a)
DEFINITION 9, \( y \geq x \). Since this argument is good for any \( G \in \mathcal{D} \), and any \( x, y \in X \), it shows that \( \delta \) is a dictator of \( \lambda G_G \).

I: Show: for any \( x, y \in X \) and \( \tilde{G}, \tilde{H} \in \mathcal{D}^{true} \): if \( \tilde{G} \mid \{x, y\} = \tilde{H} \mid \{x, y\} \), then \( f \tilde{G} \mid \{x, y\} = f \tilde{H} \mid \{x, y\} \).

Suppose \( \tilde{G} \mid \{x, y\} = \tilde{H} \mid \{x, y\} \). Choose any \( W \in [H] \). It is sufficient that there is some \( V \in [G] \) such that \( V \mid \{x, y\} = W \mid \{x, y\} \), for then since by assumption \( \lambda V \) satisfies Independence of Irrelevant Alternatives we have, by definition of \( f \) and FACT 14:

\[
 f \tilde{G} \mid \{x, y\} = {}^V \gtrless \mid \{x, y\} = {}^W \gtrless \mid \{x, y\} = f \tilde{H} \mid \{x, y\}.
\]

So let \( W \in [H] \). Consider any \( i, 1 \leq i \leq n \). Since \( \lambda G \cdot [G] \) is floating and unrafted, there is some interpretation \( \mathcal{J}_i \) such that \( W_i \in [H]^{\mathcal{J}_i} \). By our initial supposition also \( \tilde{G}_i \mid \{x, y\} = \tilde{H}_i \mid \{x, y\} \). Using LEMMA 10, choose some \( I \) and some \( V_i \in [G]^{I_i} \) such that \( V_i \mid \{x, y\} = W_i \mid \{x, y\} \). Repeating \( n \) times, construct \( V = \langle V_1, \ldots, V_n \rangle \). Plainly, \( V \mid \{x, y\} = W \mid \{x, y\} \). Since \( \lambda G \cdot [G] \) is floating and unrafted, by DEFINITION 8 we have \( V \in [G] \).

We have verified that \( f \) satisfies \( I \).

This completes the proof of THEOREM 13.

Works Cited


