Some criteria for acceptable abstraction

Øystein Linnebo
University of Bristol

Draft of November 8, 2009

An *abstraction principle* is a principle of the form

$$(\Sigma) \qquad \S\alpha = \S\beta \leftrightarrow \alpha \sim \beta$$

where the variables α and β range over entities of some sort, and where $\sim$ is an equivalence relation on this sort of entity. Frege's inconsistent Basic Law V shows that not every such principle is acceptable. A variety of criteria for acceptable abstraction have been proposed.[1]

In this note I analyze the logical relations between some of the proposed criteria, in particular conservativeness, irenicity, and stability. I answer some technical questions thrown open by the discovery of errors in the proofs of some crucial theorems in a deservedly influential study of these issues [Weir, 2003]. No attempt will be made in this note to assess the plausibility of the suggested criteria.[2]

1 Conservativeness and unboundedness

The criterion of *conservativeness* is based on a fairly intuitive idea, namely that an abstraction principle Σ is acceptable just in case it can be added to any theory without disturbing this theory's claims about the objects with which it is concerned. That is, adding Σ to a theory T may give us additional information about the 'new' objects with which Σ is concerned, but not about the 'old' objects with which T is concerned.

[1] See for instance [Boolos, 1990], [Wright, 1997], [Weir, 2003], and, for an overview, [Linnebo, 2009].

[2] See, however, [Uzquiano, 2009] and [Linnebo and Uzquiano, 2009], which challenge the status of the three criteria as respectively necessary and sufficient conditions for acceptability.

The standard way of making this intuitive idea precise is as follows. Let T be a theory in some base language $\mathcal{L}$ that does not contain the operator §. Let $\mathcal{L}^+$ be the language that results from adding to $\mathcal{L}$ the operator §. Define the predicate 'old(x)' as $\neg\exists\alpha(x = \S\alpha)$. Let ϕ^{old} be the result of restricting all the quantifiers in ϕ to 'old' objects, and let T^{old} be the result of replacing every axiom ϕ of T with ϕ^{old}.

Definition 1 (Conservativeness) An abstraction principle Σ is *conservative over an $\mathcal{L}$-theory T* iff for any $\mathcal{L}$-formula ϕ we have:

$$\text{if } T^{\text{old}} \cup \{\Sigma\} \models \phi^{\text{old}}, \text{ then } T \models \phi.$$

Σ is *conservative* iff it is conservative over any $\mathcal{L}$-theory T provided $\mathcal{L}$ does not contain the operator §.

Note that we are concerned with *semantic* rather than *proof-theoretic* conservativeness. Note also that in the definition of conservativeness there are no restrictions on the base language $\mathcal{L}$ other than that it must not contain the operator §. Henceforth this last requirement will be left implicit.

Theorem 4.2 of [Weir, 2003] asserts that any abstraction principle which is "unbounded", in the sense of the following definition, is conservative.

Definition 2 (Unboundedness) An abstraction principle Σ is *κ-satisfiable* iff Σ is satisfiable in a domain of cardinality κ. Σ is *unbounded* iff Σ is κ-satisfiable for an unbounded sequence of cardinals κ.

However, the following example shows Weir's assertion to be incorrect.

Example 1 Let Σ be the abstraction principle:

$$\S F = \S G \leftrightarrow \neg\exists x\, Px \vee \forall u(Fu \leftrightarrow Gu)$$

where F and G are monadic second-order variables and P is an atomic predicate of the base language $\mathcal{L}$. Σ is unbounded because it is satisfiable in any domain by interpreting P as not applying to any object in the domain. However, Σ is non-conservative over the theory T whose sole axiom is $\exists x\, Px$. For on the one hand we have $T^{\text{old}}, \Sigma \models \bot$ (because T^{old} ensures

that the domain contains a P, which turns Σ into Frege's inconsistent Basic Law V). But on the other hand we have $T \not\models \bot$.

What went wrong? The problem is that the notion of conservativeness requires that the vocabulary of the base language $\mathcal{L}$ retain its meaning on the 'old' domain, whereas the notion of unboundedness is defined in terms of satisfiability and thus allows this vocabulary to be reinterpreted. Any notion of unboundedness capable of implying conservativeness must ensure that the vocabulary of the base language retains its meaning on the 'old' domain. This suggests to the following definition.

Definition 3 (Uniform unboundedness) An abstraction principle Σ is *uniformly unbounded* iff for any model M of any base language $\mathcal{L}$ there is a model N of the extended language $\mathcal{L}^+$ such that:

(i) N is an extension of M whose 'old' objects of N are precisely the objects of M,

(ii) N satisfies Σ.

Lemma 1 Any uniformly unbounded abstraction principle is unbounded. But the converse does not hold.

Proof. The first claim is straightforward. The second claim follows from Example 1. $\dashv$

Theorem 1 An abstraction principle is conservative iff it is uniformly unbounded.

Proof. Assume Σ is uniformly unbounded. Assume that $T \not\models \phi$ for some $\mathcal{L}$-formula ϕ. Then there is a model M such that $M \models T \cup \{\neg\phi\}$. By the first assumption, M can be extended to a model N of Σ whose 'old' objects are precisely the objects of M. It follows that $N \models T^{\text{old}} \cup \{\Sigma, \neg\phi^{\text{old}}\}$. This shows $T^{\text{old}} \cup \{\Sigma\} \not\models \phi^{\text{old}}$. Since $\mathcal{L}$, T, and ϕ were arbitrary, it follows that Σ is conservative.

Assume next that Σ is conservative. Let M be a model of some base language $\mathcal{L}$. Let $\mathcal{L}_M$ be the enriched base language which adds to $\mathcal{L}$ a distinct constant for every element of M. ($\mathcal{L}_M$ may thus be an uncountable language, as is commonplace in model theory.) Let the $\mathcal{L}_M$-theory T consist of the diagram of M (that is, the set of all atomic sentences and negated atomic sentences in $\mathcal{L}_M$ which are true in M). Since T contains no quantifiers, we have $T^{\text{old}} = T$. Assume T, Σ has no model. Then $T^{\text{old}}, \Sigma \models \bot$, whence by Σ's

conservativeness, $T \models \bot$, which contradicts $M \models T$. So let N be a model of T, Σ. Then N can be assumed to be an extension of M whose 'old' objects are precisely the objects of M. Viewed as an $\mathcal{L}$- (rather than $\mathcal{L}_M$-) model, N then shows Σ to be uniformly unbounded. $\dashv$

Although generalized languages such as $\mathcal{L}_M$ are mathematically acceptable, it may be objected that the present use of such languages is philosophically problematic because the criterion of conservativeness was formulated with ordinary languages in mind, not generalized languages. I believe this objection is mistaken. The intuitive idea on which the criterion of conservativeness is based involves no restrictions on the base language $\mathcal{L}$ (other than that it not contain the operator §). Moreover, when we chose to study semantic rather than proof-theoretic conservativeness of higher-order theories, we already gave up on any requirement of a close link with our actual languages and deductive practices. We are in general no more able to assess questions of higher-order semantic consequence than we are to master a generalized language.[3]

I next define another assumption and show that its addition enables us to prove a partial converse of the main result of Lemma 1.

Definition 4 (Purely logical) An abstraction principle Σ is *purely logical* iff it contains no non-logical vocabulary except the operator §.

Theorem 2 Any unbounded and purely logical abstraction principle is uniformly unbounded.

Proof. Let Σ be an unboundedness and purely logical abstraction principle, and let M be a model of some base language $\mathcal{L}$. Since Σ is unbounded, it is satisfiable in a model N of cardinality larger than that of M. Since Σ is purely logical, it has no non-logical vocabulary in common with $\mathcal{L}$. (Recall that $\mathcal{L}$ has been assumed not to contain the operator §.) This ensures that N can be taken to be an extension of M whose 'old' objects are precisely those of M. $\dashv$

Corollary 1 Let Σ be a purely logical abstraction principle. Then Σ is conservative iff it is unbounded iff it is uniformly unbounded.

Proof. Immediate from Lemma 1 and Theorems 1 and 2. $\dashv$

[3] I don't know whether the theorem can be proved without any appeal to generalized languages.

2 Irenicity and stability

Unfortunately, a conservative abstraction principle need not be acceptable, as a theorem of Weir's shows ([Weir, 2003], Theorem 4.3).

Theorem 3 (Weir) There are pairs of purely logical abstraction principles each of which is conservative but which are not jointly satisfiable.

A key ingredient of the proof of Theorem 3 is the following, well-known theorem.

Theorem 4 (Folklore) In the language of pure second-order logic we can characterize various cardinality properties of a concept X, such as: being of size $\aleph_n$ for some natural number n, being of continuum size, being of limit-cardinal size, being of successor-cardinal size, and being of inaccessible size.

Proof. See for instance [Shapiro, 2000], pp. 104-105. ⊣

Proof of Theorem 3. Consider the following restricted version of Frege's Basic Law V:

(RV) $$\epsilon F = \epsilon G \leftrightarrow (\textsc{Bad}(F) \wedge \textsc{Bad}(G)) \vee \forall x(Fx \leftrightarrow Gx)$$

where $\textsc{Bad}(F)$ is some $\mathcal{L}$-formula. By Theorem 4 we can let $\textsc{Bad}_1(F)$ and $\textsc{Bad}_2(F)$ express that the universal concept U is respectively of successor-cardinal size and limit-cardinal size. The resulting versions of (RV) are easily seen to be satisfiable in all and only domains of respectively successor-cardinal size and limit-cardinal size. So the two principles are not jointly satisfiable. But each is unbounded and thus conservative by Corollary 1. ⊣

The next definition is a natural response to the problem posed by Theorem 3.

Definition 5 (Stability) An abstraction Σ is *stable* iff there is a cardinal κ such that Σ is λ-satisfiable for all cardinals $\lambda \geq \kappa$. Σ is *strongly stable* iff there is a cardinal κ such that Σ is λ-satisfiable just in case $\lambda \geq \kappa$.

Theorem 6.1 of [Weir, 2003] asserts that stability is equivalent to another criterion that has been proposed, namely irenicity.

Definition 6 (Irenicity) An abstraction Σ is *irenic* iff it is conservative and jointly satisfiable with any other conservative abstraction principle.

However, Weir's proof is flawed for two independent reasons. The first flaw is brought out by Example 1, which shows that an abstraction principle can be strongly stable without being conservative and thus *a fortiori* without being irenic. Corollary 1 suggests that this problem can be avoided by adding the assumption that Σ is purely logical, which will be confirmed below. But even with this added assumption, an observation due to Stewart Shapiro shows the proof to contain a second, independent flaw, namely a failure to distinguish properly between stability and strong stability. All Weir's proof establishes is that (assuming pure logicality) strong stability entails irenicity, which in turn entails stability. The proof thus leaves open the question whether any of the converses hold. This will now be investigated.

Lemma 2 Let Σ be an abstraction principle that is not stable. Then there is a conservative and purely logical abstraction principle Γ which is not jointly satisfiable with Σ.

Proof. By Corollary 1 it suffices to find a purely logical abstraction principle Γ that is satisfiable at precisely those cardinalities where Σ is not satisfiable. I claim that it is possible to formulate a condition $\textsc{Bad}(F)$ which expresses that Σ isn't satisfiable at the universal concept U. If so, the resulting version of (RV) will fit our bill. To see this, assume first that Σ isn't κ-satisfiable. Then all concepts on a domain of size κ are Bad, which makes Γ trivially κ-satisfiable. Next, assume that Σ is κ-satisfiable. Then no concepts on a domain of size κ are Bad, which means that on such domains Γ is like Basic Law V and thus not κ-satisfiable.

It remains to prove the claim. Let $\mathcal{R}(\Sigma)$ be the Ramseyfication of Σ, which is available in a language of order one higher than that of Σ. Then $\textsc{Bad}(F)$ can be chosen to be $\neg\mathcal{R}(\Sigma)$.

(In fact, the claim can, if desired, be proved without having to ascend to a language of order higher than Σ. Consider the case where Σ is second-order; other cases are analogous. Then a dyadic relation R can be used to code an assignment of objects to selected monadic concepts by letting $\forall u(Fu \leftrightarrow Rux)$ mean that x is assigned to the concept F. Let $\sim$ be the equivalence relation on which Σ abstracts. Say that F *is associated with* x *under* R iff F bears $\sim$ to some concept F' which R associates with x. The claim that Σ is satisfiable can then be expressed as the claim that there is a dyadic relation R such that every concept F is associated with an object x under R, and two concepts F and G are associated with the

same object under R just in case $F \sim G$.) $\dashv$

Theorem 5 Any irenic abstraction principle is stable. But the converse does not hold.

Proof. Assume Σ is not stable. Then by Lemma 2 there is a conservative abstraction principle Γ with which Σ is not jointly satisfiable, which shows that Σ is not irenic. Example 1 shows that the converse does not hold. $\dashv$

However, as in the case of Theorem 2, a converse holds under the added assumption of pure logicality.

Theorem 6 Assume Σ is purely logical and stable. Then Σ is irenic.

Proof. Assume Σ is purely logical and stable. Since stability implies unboundedness, Σ is conservative by Corollary 1. Let Γ be another conservative abstraction principle. We need to show that Σ and Γ are jointly satisfiable. Let κ be a cardinal such that Σ is satisfiable at any domain of cardinality $\geq \kappa$. But by Theorem 1 and Lemma 1 there is a cardinal $\lambda \geq \kappa$ such that Γ too is satisfiable at domains of cardinality λ. It follows that Σ is irenic. $\dashv$

Corollary 2 There are purely logical and irenic abstraction principles that are not strongly stable.

Proof. This is immediate from Theorem 6 and the observation that there are purely logical abstraction principles that are stable but not strongly stable. To establish this observation, consider for instance the version of (RV) where the condition $\textsc{Bad}(F)$ is defined so as to be true in domains of cardinality $\aleph_0$ and $\geq \aleph_\omega$ but not in domain of any other cardinality. This condition can be expressed by Theorem 4. $\dashv$

References

[Boolos, 1990] Boolos, G. (1990). The Standard of Equality of Numbers. In Boolos, G., editor, *Meaning and Method: Essays in Honor of Hilary Putnam*. Harvard University Press, Cambridge, MA. Reprinted in [Boolos, 1998].

[Boolos, 1998] Boolos, G. (1998). *Logic, Logic, and Logic*. Harvard University Press, Cambridge, MA.

[Hale and Wright, 2001] Hale, B. and Wright, C. (2001). *Reason's Proper Study*. Clarendon, Oxford.

[Linnebo, 2009] Linnebo, Ø. (2009). Introduction [to a special issue on the bad company problem]. *Synthese*, 170(3):321–329.

[Linnebo and Uzquiano, 2009] Linnebo, Ø. and Uzquiano, G. (2009). Which Abstraction Principles Are Acceptable? Some Limitative Results. *British Journal for Philosophy of Science*, 60(2):239–253.

[Shapiro, 2000] Shapiro, S. (2000). *Foundations without Foundationalism: A Case for Second-Order Logic*. Oxford University Press, Oxford.

[Uzquiano, 2009] Uzquiano, G. (2009). Bad company generalized. *Synthese*, 170(3):331–347.

[Weir, 2003] Weir, A. (2003). Neo-Fregeanism: An Embarrassment of Riches. *Notre Dame Journal of Formal Logic*, 44:13–48.

[Wright, 1997] Wright, C. (1997). The Philosophical Significance of Frege's Theorem. In Heck, R., editor, *Language, Thought, and Logic. Essays in Honour of Michael Dummett*. Clarendon, Oxford. Reprinted in [Hale and Wright, 2001].