Journal of Philosophical Logic (2019) 48:867–884
https://doi.org/10.1007/s10992-018-09499-0

The Harmony of Identity

Ansten Klev[1]

Received: 4 November 2018 / Accepted: 27 December 2018 / Published online: 12 January 2019

Abstract
The standard natural deduction rules for the identity predicate have seemed to some not to be harmonious. Stephen Read has suggested an alternative introduction rule that restores harmony but presupposes second-order logic. Here it will be shown that the standard rules are in fact harmonious. To this end, natural deduction will be enriched with a theory of definitional identity. This leads to a novel conception of canonical derivation, on the basis of which the identity elimination rule can be justified in a proof-theoretical manner.

Keywords Identity · Natural deduction · Definition

1 Introduction

Identity, as is well-known, "gives rise to challenging questions which are not altogether easy to answer." One such question, discussed in some recent literature, is how to provide the identity predicate, =, of predicate logic with harmonious natural deduction rules. Introduction and elimination rules for the identity predicate were provided by logicians long ago, but it is not obvious that these rules are harmonious; more precisely, it is not obvious that the introduction rule is strong enough to justify the elimination rule. Stephen Read [16] proposed an alternative introduction rule that he showed to justify the elimination rule. One may, however, ask whether Read's rule is not too strong, since—as we shall see below—it seems to require second-order logic for its proper functioning. Here it will be shown that the standard rules for identity are in fact harmonious, hence that we do not need Read's revised introduction rule.

When justifying a rule from a proof-theoretical point of view, one must take certain rules or derivations as valid outright. Such rules or derivations are usually called canonical. It is common to assign the office of canonical derivation to derivations

✉ Ansten Klev
klev@flu.cas.cz

[1] Institute of Philosophy, Czech Academy of Sciences, Prague, Czech Republic

that end in the application of an introduction rule. This stipulation is motivated by the thought that introduction rules determine the meaning of the formula-forming operators (connectives, quantifiers, predicates). In showing the harmony of the standard rules of identity I shall rely on a more general conception of canonical derivation: it is a derivation that ends in the application of an introduction rule followed by any number, possibly zero, of substitutions of terms or formulae for definitionally identical terms or formula. This is a reasonable generalization of the standard conception, since such a sequence of substitutions is just a rewriting of the end formula of the derivation.

The notion of definitional identity invoked here will be explained by laying down certain principles it is taken to satisfy. More specifically, I shall describe a formal system whose formulae are all of the form $a \equiv b$, expressing that a and b are definitionally identical. This formal system can be integrated into natural deduction via a rule allowing the substitution of b for a definitionally identical a in a derived formula A. There is in fact independent motivation for introducing definitional identity into natural deduction once the identity predicate is present. Definitions form an important source of theorems of the form $t = u$, for syntactically distinct t and u. Hence, when identity is present, definitions should somehow be accounted for in the formalism. Formulating definitions in terms of the ordinary identity predicate forces one, however, to give up the topic-neutrality of identity. If definitions are formulated in terms of a separate notion of definitional identity, by contrast, the topic-neutrality of identity can be preserved. Once a theory of definitional identity is in place, moreover, one can see that the relation of definitional identity is strictly finer than the relation determined by the identity predicate. Definitional identity is thus not already contained in the logic of ordinary identity formulae $t = u$. A subsidiary aim of what follows is to draw attention to the importance to the logic of identity of the notion of definitional identity.

In Section 2 the standard rules for the identity predicate are presented, and it is explained precisely what is taken to be the problem with them. Read's alternative introduction rule is discussed in Section 3. Definitional identity is introduced in Section 4 and given a more formal treatment in Section 5. The new conception of canonical derivation is presented in Section 6. Against this background it is shown in the final Section 7 that the standard rules for identity are in fact in harmony.

2 The Problem

I shall assume the elegant set of natural deduction rules for identity formulated by Martin-Löf [9, p. 190]:

$$t = t \qquad \text{(=-INTRO)}$$

$$\frac{t = u \qquad A[x, x]}{A[t, u]} \qquad \text{(=-ELIM)}$$

Here t and u are arbitrary terms of the language, and the formula $A[x, x]$ arises from $A[y, z]$ by substituting the variable x for both of the free variables y and z. An application of =-ELIM binds the variable x in the derivation above the conclusion $A[t, u]$. A derivation ending in $A[x, x]$ may be thought of as showing that A defines a reflexive relation on the underlying domain, D. The rule =-ELIM in effect says that identity is the smallest reflexive relation on D. An alternative elimination rule for identity [3, 8], a form of the indiscernibility of identicals,

$$\frac{t = u \qquad B[t]}{B[u]} \tag{Ind-Id}$$

is equivalent to =-ELIM in the presence of the rules for implication and the universal quantifier.

A derivation ending in an application of =-INTRO followed by =-ELIM can be reduced as follows:

$$\frac{t = t \qquad \begin{matrix}\mathscr{D}[x] \\ A[x, x]\end{matrix}}{A[t, t]} \qquad \rightsquigarrow \qquad \begin{matrix}\mathscr{D}[t] \\ A[t, t]\end{matrix}$$

Indeed, a normalization theorem can be proved for first-order logic extended with a general scheme of rules for so-called inductively defined predicates, of which identity as captured by =-INTRO and =-ELIM is an instance [9].

In spite of this mathematical result, several authors have found that the proposed rules are not in harmony [5, 12, 16, 17]. The cause of the felt disharmony is the difference between the form of the conclusion of =-INTRO, viz. $t = t$, and the form of the major premiss of =-ELIM, viz. $t = u$. This difference in form makes it unclear how one should go about justifying =-ELIM on the basis of =-INTRO.

The justification of an elimination rule on the basis of introduction rules can be carried out more or less formally. Let us first consider the less formal way. An elimination rule for an operator Φ may be written schematically as follows:

$$\frac{B \qquad \text{minor premisses}}{C}$$

Here B is a formula whose outermost operator is Φ. The minor premisses may be formulae or whole derivations. To justify the rule informally we may proceed as follows:

1. Assume that the major premiss B is the conclusion of an application of Φ-INTRO.
2. Use the premisses of this application of Φ-INTRO together with the minor premisses in order to justify the conclusion C.

The rationale behind this procedure is the stipulation—first made by Gentzen [4]—that the Φ-INTRO rules determine the meaning of Φ. Given this stipulation, the assumption that B is the conclusion of an application of Φ-INTRO amounts to the assumption that we have analyzed the meaning of B. In step 2 we then use this analysis together with the minor premisses in justifying the conclusion C.

As an example we may consider $\supset$-ELIM:

$$\frac{A \supset C \qquad A}{C}$$

Assume that the major premiss, $A \supset C$, is the conclusion of an application of $\supset$-INTRO:

$$\dfrac{\begin{array}{c}[A]\\ \vdots\\ C\end{array}}{A \supset C}$$

Then we have a derivation $\mathscr{D}_1$ of C from A as open assumption. The minor premiss of $\supset$-ELIM gives a derivation $\mathscr{D}_2$ of A. But then we have a derivation of C as follows:

$$\begin{array}{c}\mathscr{D}_2\\ A\\ \mathscr{D}_1\\ C\end{array}$$

The same procedure can be carried out for the rest of the familiar logical connectives and the quantifiers. It cannot, however, be used to justify =-ELIM on the basis of =-INTRO, for at least two reasons. Firstly, owing to the difference in form between the major premiss of =-ELIM and the conclusion of =-INTRO, we cannot assume that such a major premiss is the conclusion of an application of =-INTRO. Secondly, even in a case where the major premiss of =-ELIM has the form $t = t$ and may be assumed to be the conclusion of an application of =-INTRO, this introduction rule has no premisses that can be used in step 2.

A more formal way of justifying an elimination rule on the basis of introduction rules relies on Prawitz's notion of the validity of a derivation [14, 15, 18]. Assume that we have a notion of validity of derivations. For our purposes it is enough to consider rules of the form

$$\dfrac{A_1 \qquad \dots \qquad A_n}{C}$$

Such a rule can then be said to be justified if, whenever the derivations

$$\begin{array}{c}\mathscr{D}_1\\ A_1\end{array} \quad \dots \quad \begin{array}{c}\mathscr{D}_n\\ A_n\end{array}$$

are valid, then so is the derivation

$$\dfrac{\begin{array}{c}\mathscr{D}_1\\ A_1\end{array} \qquad \dots \qquad \begin{array}{c}\mathscr{D}_n\\ A_n\end{array}}{C}$$

A derivation is here understood as a tree of formulae together with information regarding the discharging of open assumptions and the binding of free variables. That a derivation $\mathscr{D}$ is closed means that all of its assumptions have been discharged and all of its free variables have been bound.

Certain closed derivations are to be taken as valid outright (provided all immediate sub-derivations are valid). These derivations are called *canonical*. That we recognize a notion of canonical derivation is a reflection of the fact that certain rules are taken to determine the meaning of the formula-forming operators. In particular, where—as here—introduction rules are taken to be meaning-determining, a canonical derivation is a derivation that ends in the application of an introduction rule. The term 'canonical derivation' thus designates a certain office that (for the time being) we take to be filled by derivations ending in an introduction rule. And we say that a canonical derivation is valid iff all its immediate sub-derivations are valid.

A closed derivation that is not canonical is valid iff it can be *reduced* to a valid canonical derivation. The notion of reduction assumed here is to include the reductions introduced by Prawitz [13], namely transformations of derivations such as

$$\dfrac{\dfrac{\begin{matrix}A\\ \mathscr{D}_1\\ C\end{matrix}}{A \supset C} \qquad \begin{matrix}\mathscr{D}_2\\ A\end{matrix}}{C} \qquad \rightsquigarrow \qquad \begin{matrix}\mathscr{D}_2\\ A\\ \mathscr{D}_1\\ C\end{matrix}$$

A first attempt at defining reduction in general is by saying that a reduction of a derivation $\mathscr{D}$ is a transformation of $\mathscr{D}$ that preserves the conclusion of $\mathscr{D}$ and that does not introduce any new open assumptions or free variables. It is, however, natural to require something much stronger of a reduction, namely that it preserves the identity of derivations. Without this stronger requirement it is difficult to see the motivation behind the stipulation that a closed non-canonical derivation is valid iff it can be reduced to a valid canonical derivation. With the requirement, by contrast, the stipulation says that a closed derivation is valid iff it is identical to a valid canonical derivation, and that clearly seems like a reasonable stipulation. The reductions we shall consider in this paper can all be seen to meet the stronger requirement.

The validity of an open derivation $\mathscr{D}$ is explained in terms of the validity of the closed derivations that arise from $\mathscr{D}$ by substituting closed terms for variables free in $\mathscr{D}$ and supplying the open assumptions in $\mathscr{D}$ with valid canonical derivations. The motivation behind this stipulation is that an open assumption A in a derivation stands for an arbitrary derivation of A and a free variable stands for an arbitrary term. More precisely, we make the following stipulations.

Suppose the variable x is free in $\mathscr{D}$. We may write the derivation as $\mathscr{D}[x]$. This derivation is defined to be valid iff $\mathscr{D}[t]$ is valid whenever t is a closed term.

Suppose the derivation $\mathscr{D}$ contains an open assumption A. We may write the derivation as

$$\begin{matrix}A\\ \mathscr{D}\end{matrix}$$

This derivation is defined to be valid iff for any valid canonical derivation $\mathscr{D}'$ whose end formula is A, the derivation

$$\begin{matrix}\mathscr{D}'\\ A\\ \mathscr{D}\end{matrix}$$

is valid.

Again we may consider the example of $\supset$-ELIM. Assume that we have valid derivations

$$\begin{matrix}\mathscr{D}_1\\ A \supset B\end{matrix} \qquad \begin{matrix}\mathscr{D}_2\\ A\end{matrix}$$

We must show that the derivation

$$\dfrac{\begin{matrix}\mathscr{D}_1\\ A \supset B\end{matrix} \qquad \begin{matrix}\mathscr{D}_2\\ A\end{matrix}}{B} \tag{$\supset$-ElimDer}$$

is valid. If $\supset$-ElimDer is open, then we substitute closed terms for free variables and supply open assumptions with valid canonical derivations. The resulting immediate subderivations $\mathscr{D}_1'$ and $\mathscr{D}_2'$ will then be closed, since $\supset$-ELIM neither discharges any open assumptions nor binds any variable. Hence we may assume that $\mathscr{D}_1$ and $\mathscr{D}_2$ are already closed. Since they are also valid, they reduce to valid canonical derivations. Carrying out these reductions on $\supset$-ElimDer yields a derivation

$$\dfrac{\dfrac{\begin{array}{c}A\\ \mathscr{D}_1^*\\ C\end{array}}{A \supset C} \qquad \begin{array}{c}\mathscr{D}_2^*\\ A\end{array}}{C} \qquad (\supset\text{-ElimDer*})$$

Here $\mathscr{D}_1^*$ is a valid derivation whose only open assumption is A, and $\mathscr{D}_2^*$ is a valid canonical derivation. On $\supset$-ElimDer* we may carry out $\supset$-reduction to get a derivation

$$\begin{array}{c}\mathscr{D}_2^*\\ A\\ \mathscr{D}_1^*\\ C\end{array}$$

Since $\mathscr{D}_1^*$ is valid and its only open assumption is A, any result of supplying this assumption with a valid canonical derivation of A is itself valid. But $\mathscr{D}_2^*$ is a valid canonical derivation, hence the displayed derivation is valid. Since it is also closed, it reduces to a valid canonical derivation. By the transitivity of reduction, also $\supset$-ElimDer reduces to a valid canonical derivation, hence it is itself valid.

Again one can see that this mode of justification can be carried out for all the connectives and quantifiers, but not for =-ELIM. Namely, let us assume that we have valid derivations

$$\begin{array}{c}\mathscr{D}_1\\ t = u\end{array} \qquad \begin{array}{c}\mathscr{D}_2\\ A[x, x]\end{array}$$

and try to show that the derivation

$$\dfrac{\begin{array}{c}\mathscr{D}_1\\ t = u\end{array} \qquad \begin{array}{c}\mathscr{D}_2\\ A[x, x]\end{array}}{A[t, u]} \qquad (=\text{-ElimDer})$$

is valid. As in the case of $\supset$-ELIM, we may assume that the derivation =-ElimDer is closed. The subderivation $\mathscr{D}_1$ is therefore closed and valid, hence it reduces to a valid canonical derivation

$$\begin{array}{c}\mathscr{D}_1'\\ t = u\end{array}$$

At this point, however, we have reached a dead end, since we have not specified what a canonical derivation looks like for the formula $t = u$ where t and u are syntactically different terms. Our stipulation that a canonical derivation is one that ends in an introduction rule together with the rule =-INTRO serve to specify only what a canonical derivation of $t = t$ looks like.

Since a closed valid derivation need not end in an introduction rule, we are not entitled to conclude from this limitation of =-INTRO that there can be no closed valid derivation of $t = u$ for syntactically distinct t and u. Instead, we must conclude that we have not provided enough information as to what a canonical derivation of an identity formula looks like. Lacking such information, we in effect do not know what it means for the closed derivation $\mathscr{D}_1$ above to be valid. We are therefore stuck in our attempt to justify =-ELIM, since we have reached a thesis the meaning of which we have not specified fully.

3 Read's Rule

In response to this problem, Read [16] has suggested that we change the introduction rule for the identity predicate to the following:

$$\dfrac{\begin{array}{c}[F(t)]\\ \vdots\\ F(u)\end{array}}{t = u} \qquad \text{(=-INTRO*)}$$

Here F is a unary predicate variable that does not occur free in any assumption other than $F(t)$, which assumption is discharged by an application of =-INTRO*. Read's starting point is that the indiscernibility of identicals,

$$\dfrac{t = u \qquad B[t]}{B[u]} \qquad \text{(Ind-Id)}$$

should be captured by the elimination rule for identity. The introduction rule should, accordingly, capture the identity of indiscernibles, as =-INTRO* indeed seems to do. If we instead take the rule =-ELIM as our elimination rule, then the introduction rule analogous to =-INTRO* is:

$$\dfrac{\begin{array}{c}[R(x, x)]\\ \vdots\\ R(t, u)\end{array}}{t = u}$$

Here R is a binary predicate variable and x an individual variable, neither of which occurs free in any assumption other than $R(x, x)$. An application of the rule discharges the assumption $R(x, x)$ and binds the variable x above the conclusion $t = u$. The remarks that follow pertaining to =-INTRO* apply equally well to this rule.

The elimination rule that Read takes to correspond to =-INTRO* is not the full indiscernibility of identicals, Ind-Id, but rather the following restricted version of it:

$$\dfrac{t = u \qquad F(t)}{F(u)} \qquad \text{(=-ELIM*)}$$

As before, F is a predicate variable, hence $F(t)$ is not a schematic formula such as the minor premiss and conclusion in Ind-Id are. It is, however, clear that in order to have the right to call =-INTRO* and =-ELIM* rules of identity, we need to show that Ind-Id follows from =-ELIM*. It is, moreover, clear that in order to show this, we need rules governing the predicate variable F. Such rules are, however, available only in second-order logic. Thus, Read [16, p. 117] invokes rules for the second-order existential quantifier when showing that Ind-Id follows from =-ELIM*. An argument not invoking second-order logic is offered by Read [17, p. 416], but this argument fails, it seems to me. It proceeds by induction on the complexity of the minor premiss $B[t]$ of Ind-Id. The induction steps as given by Read seem to be in order, but the base case is dealt with too quickly. The base case, where $B[t]$ is atomic, is said to follow from =-ELIM*. But, in the absence of rules governing predicate variables, we can apply =-ELIM* to get $B[u]$ from $B[t]$ and $t = u$ only if $B[t]$ is of the form $F(t)$, for a predicate variable F; and there might well be atomic formulae that are not of this form.

In order to apply =-INTRO* to get $t = u$, for syntactically distinct t and u, we need a derivation of $F(u)$ from $F(t)$. Also such a derivation requires rules governing predicate variables. For without any such rules, $F(u)$ and $F(t)$ are in effect just distinct propositional variables; and it is immediate from normalization that in this case there is no derivation of $F(u)$ from $\{F(t)\} \cup \Gamma$ unless $F(u) \in \Gamma$; but if $F(u) \in \Gamma$, then =-INTRO* is not applicable owing to the restriction on occurrences of F in open assumptions.[1]

The rule =-INTRO* does therefore not seem to be an option for those who want to avoid second-order logic. For someone who is not averse to second-order logic, and who may therefore be willing to accept =-INTRO* as the introduction rule for identity, it is, however, not clear why separate rules should be given for the identity predicate: in second-order logic the identity predicate can be explicitly defined, and an expression that is so defined has its meaning completely determined by its definition.

4 A Role for Definitions

Suppose that in light of these difficulties pertaining to Read's rule =-INTRO* we decide to hold on to =-INTRO. Our problem is then to explain what a canonical derivation of $t = u$, for syntactically distinct terms t and u, looks like. Our stipulations so far fail to tell us this. The relevant stipulations are:

(i) A canonical derivation is one that ends in the application of an introduction rule.
(ii) The introduction rule for identity is =-INTRO.

Read's suggestion is to change stipulation (ii). I suggest that we instead look at stipulation (i). Thus I wish to propose that we revise the stipulation of what is to count as a canonical derivation. I will say that a canonical derivation is one that ends in the application of an introduction rule followed by any number, possibly zero, of replacements of terms or formulae by definitionally identical terms or formulae. This is a natural generalization of stipulation (i) once the relation of definitional identity is present, since the substitution of an expression a for a definitionally identical b in an expression c is just a rewriting of c. A canonical derivation in the new sense is therefore just a canonical derivation in the old sense possibly followed by certain rewritings of the conclusion. That the relation of definitional identity should be included in the

[1] This result is proved model-theoretically in Milne [12, p. 38, fn. 12]; it is the main topic of Griffiths [5].

formalism once the identity predicate is present can be seen by reflections such as the following.

Many forms of definition take the form of one or more equations; and on the basis of such definitions we can derive formulae of the form $t = u$ for syntactically different terms t and u. For instance, from the well-known recursive definition of addition and the definition of 1 as $\mathbf{s}(0)$, the successor of 0, we can derive

$$1 + 1 = \mathbf{s}(\mathbf{s}(0))$$

and infinitely many other identity formulae. In trying to understand the logic of identity we are therefore led to ask how definitions are to be dealt with in a formal system. We may distinguish two kinds of definition (we are interested only in definitions that take the form of equations). A definition of the first kind introduces vocabulary that is to abbreviate an expression already available in the language; this is often called an explicit definition. In this case the *definiendum* is to be eliminable, and the extension of the theory by the definition is to be conservative (i.e., the definition is not "creative"). A definition of the second kind, by contrast, is to determine the meaning of primitive vocabulary. Here the *definiendum* need not be eliminable, nor need a theory extended by such a definition be conservative. Such definitions therefore have to appear as axioms, such as the defining equations of addition and multiplication do in first-order Peano arithmetic. It is, however, common also to regard an explicit definition as an axiom of an extended (conservatively extended) theory. Definitions quite generally are thus typically treated as axioms in formal systems.

It is natural to define the notion of an axiom as a rule without premisses. Accordingly, we may regard =-INTRO as an axiom; and, conversely, any further axiom of the form $t = u$ that a theory contains may be regarded as an introduction rule for the identity predicate. Indeed, such an axiom serves, just as much as =-INTRO does, to introduce the identity predicate into a derivation. If definitions take the form $t = u$ and are regarded as axioms, they must therefore also be regarded as introduction rules for the identity predicate. Since introduction rules are meaning-determining, definitions thus come to play a role in determining the meaning of the identity predicate.

The account of identity that this leads to is, however, quite unsatisfactory. Behind the usual treatment of the identity predicate as a logical symbol lies the idea that the rules governing it are topic neutral: they are the same regardless of topic, that is, the same in any domain of discourse. But if defining equations are introduction rules, then the identity predicate can no longer be regarded as topic neutral, since its rules will then differ from one theory to the other. The defining equations needed in, for instance, arithmetic are unlike those needed in set theory, hence these two theories will provide the identity predicate with different introduction rules, and so with a different meaning. In fact, not only shall we have to regard the meaning of identity in arithmetic as being different from its meaning in set theory: even the simple addition of an explicit definition to a theory will have to be regarded as changing the meaning of the identity predicate inside that theory, since such a definition will be a new introduction rule for the identity predicate.

Adding a theory of definitional identity to natural deduction will not only make it possible to justify =-ELIM on the basis of =-INTRO. It will also make it clear that definitions do not disturb the topic-neutrality of the identity predicate.

5 Definitional Identity

Accounts of definitional identity have already been developed by Curry and Feys [2, pp. 62–76], Curry [1, pp. 104–111], and Martin-Löf [10], and we may take these accounts as our starting point. The formulae of the theory are all of the form $a \equiv b$, expressing that a and b are definitionally identical. Thus we have an equational theory (no connectives or quantifiers!); but a and b may be expressions of any category, in particular, terms or formulae, provided they are both of one and the same category. The principles governing definitional identity will depend on the underlying language, in particular on which categories of expression are available and on how composite expressions may be formed. For the language of first-order predicate logic, the following seems like a reasonable characterization.

Definitional identity is an equivalence relation, thus there are first of all the following axiom and rules:

$$a \equiv a \qquad \frac{a \equiv b}{b \equiv a} \qquad \frac{a \equiv b \quad b \equiv c}{a \equiv c}$$

In addition there are two axiom schemes and two rules, as well as a “bridge principle” that lets us integrate a derivation $\mathfrak{D}$ from the theory of definitional identity into an ordinary natural deduction derivation $\mathscr{D}$.

The theory of definitional identity is understood to be paired with an underlying theory. We assume that for the theory in question specifications have been made regarding the formal conditions that a definition has to meet. The first axiom scheme says that whenever a and b meet these conditions as *definiendum* and *definiens* respectively, then $a \equiv b$ may be posited as an axiom. In schematic form we may write:

$$\textit{definiendum} \equiv \textit{definiens}$$

In first-order arithmetic, for instance, the modes of definition usually admitted are explicit definition of individual constants and of functions, as well as recursive definition of functions. An explicit definition, of an individual constant c or a function f, takes the form of a simple equation:

$$c \equiv t \qquad\qquad f(\bar{x}) \equiv t[\bar{x}]$$

A recursive definition consists of two equations:

$$\begin{aligned} &f(\bar{x}, 0) \equiv t[\bar{x}] \\ &f(\bar{x}, \mathbf{s}(y)) \equiv u[\bar{x}, y, f(\bar{x}, y)] \end{aligned}$$

When adding a recursive definition to an underlying arithmetical theory, we are thus positing two axioms in the theory of definitional identity.[2]

It seems reasonable to require that definitional identity be preserved by the renaming of bound variables, since the result of such a renaming is just a notational variant of what one started out with. The second axiom scheme says that a formula A is definitionally identical to any formula B that arises from A by the renaming of one or more of its bound variables:

$$A \equiv B$$

Of course, we must take care that B is really just a notational variant of A, that no new binding relations are created in B. In standard first-order logic, the only variable-binding operators are the existential and the universal quantifiers, each of which produces a formula. If there are variable-binding operators producing expressions of other grammatical categories, for instance a definite-description operator, then a similar rule is posited for each such category.

When we analyze the meaning of an expression by continuously replacing in it *definienda* by their *definientia*, then it seems clear that definitional identity is preserved. For instance, if 2 is definitionally identical to $\mathbf{s}(1)$, and 1 is definitionally identical to $\mathbf{s}(0)$, then it seems clear that 2 is definitionally identical to $\mathbf{s}(\mathbf{s}(0))$. The general principle is captured by the following rule:

$$\frac{a \equiv b \qquad c \equiv c'}{a \equiv b[c'/c]_!} \tag{R1}$$

Here $b[c'/c]_!$ is any formula that results from replacing any number of occurrences of c in b by c'.[3] It is of course presupposed that a is of the same category as b, and that c is of the same category as c'; but c may, for instance, be a term, and a a formula.

In order to be able to state the definition of a function, such as the definition of addition, by means of variables, we need a way of instantiating terms for variables. Clearly, the instantiation should preserve definitional identity. For instance, given the definitional equation $x + 0 \equiv x$, it should follow that $2 + 0 \equiv 2$, i.e., that $2 + 0$ is definitionally identical to 2. Thus we have the following rule:

$$\frac{a \equiv b}{a[t/x] \equiv b[t/x]} \tag{R2}$$

[2] For the benefit of conceptual clarity we might distinguish recursive definition from definition by cases. The standard definition of the predecessor function,

$$\begin{gathered}\mathbf{pred}(0) \equiv 0 \\ \mathbf{pred}(\mathbf{s}(y)) \equiv y\end{gathered}$$

is a definition by cases, but it is not recursive, that is, it does not refer to "previous" values of itself. The displayed scheme of recursive definition in arithmetic covers also such non-recursive definitions by cases.

[3] Unlike the ordinary substitution notation, the notation $b[c'/c]_!$ therefore does not uniquely determine the formula in question. If we need $b[c'/c]_!$ in a given context to stand for different formulae (namely, formulae that arise by substituting c' for different sets of occurrences of c in b), then we may use some mechanism such as priming for indicating this. Otherwise it will be assumed that $b[c'/c]_!$ at all of its occurrences stands for one and the same formula.

Here x is a first-order variable, t is a term, and $a[t/x]$ is the result of replacing x by t at all of its occurrences in a. If variables of other types are available, then a similar rule is posited for each type.

In arithmetic the schemes of explicit definition and definition by recursion together with the rules R1 and R2 allow us to define all primitive recursive functions [6, § 54]. Indeed, the theory of definitional identity quite generally may be regarded as a theory of formal computation: the unravelling of definitions inside an expression, as can be carried out within a theory of definitional identity, may naturally be thought of as its calculation, computation, or evaluation. The value of an expression is then its complete analysis. By incorporating definitional identity into natural deduction, we thus get a system that formalizes not only proof, but also computation.

As the "bridge principle" that lets us integrate derivations $\mathfrak{D}$ from the theory of definitional identity into ordinary natural deduction derivations we shall make use of the following rule:

$$\frac{A \qquad a \equiv b}{A[b/a]_!}$$

We shall call this rule *definitional substitution*. Notice that the conclusion $A[b/a]_!$ of an application of definitional substitution is definitionally identical to, hence just a rewriting of, the left premiss A.

Consider a derivation $\mathscr{D}$ whose final step is a definitional substitution:

$$\frac{\begin{matrix}\mathscr{D}' \\ A\end{matrix} \qquad \begin{matrix}\mathfrak{D} \\ a \equiv b\end{matrix}}{A[b/a]_!}$$

All top formulae of $\mathfrak{D}$ count as axioms in $\mathscr{D}$. That is, they are not assumptions that may be discharged in any extension of $\mathscr{D}$. We might therefore think of $\mathfrak{D}$ as a separate derivation and picture $\mathscr{D}$ as follows:[4]

$$\begin{array}{c}\mathscr{D}' \\ A \\ \hdashline A[b/a]_!\end{array}\ \mathfrak{D}$$

The derivation $\mathfrak{D}$ has been written down on a separate sheet of paper, say, and is invoked here only in order to justify the rewriting of A, the end formula $\mathscr{D}'$. A dashed line is used to indicate that this final step in $\mathscr{D}$ is just such a rewriting. This way of regarding the bridge principle leads us to stipulate that $\mathscr{D}$ has only one immediate sub-derivation, namely $\mathscr{D}'$ (including its end formula A). Thus, we shall not count $\mathfrak{D}$ as a sub-derivation of $\mathscr{D}$.

An alternative, but equivalent, bridge principle is:

$$\frac{A \equiv B}{A \leftrightarrow B}$$

Here $A \leftrightarrow B$ is ordinary material equivalence, hence the rule says that the equivalence relation of definitional identity among formulae is at least as fine as material equivalence.

[4] Cf. the rule of formula conversion in [11, p. 155].

As a special case of definitional substitution we have the rule that allows the renaming of bound variables. Let A and B differ at most in the name of their bound variables. Then we have

$$\frac{A \qquad A \equiv B}{B}$$

as a special case of definitional substitution. Borrowing a well-known terminology from Curry and Feys [2, p. 90], the derived rule that has A as its only premiss and such a B as conclusion will be called α.

A bridge principle for the category of terms,

$$\frac{t \equiv u}{t = u}$$

is easily derivable by means of =-INTRO and definitional substitution. This rule says that the equivalence relation of definitional identity among terms is at least as fine as the relation determined by the identity predicate. In arithmetic it can be shown that definitional identity is in fact strictly finer than the relation determined by the identity predicate. Namely, although the formula $x + y = y + x$, for variables x and y, is derivable by (quantifier-free) induction, the corresponding definitional identity $x + y \equiv y + x$ is not derivable. Definitional identity in first-order arithmetic can be seen to be the equivalence relation generated by a reduction relation for which strong normalization and a Church–Rosser theorem can be proved.[5] Since both $x + y$ and $y + x$, for variables x and y, are irreducible but not syntactically identical terms, it then follows that they are not definitionally identical. This shows that definitional identity is a non-trivial addition to ordinary predicate logic.

6 Canonical Derivations

We shall take a canonical derivation to be a derivation that ends in an introduction rule followed by any number, possibly zero, of definitional substitutions. An operator Φ may apply to formulae, to terms, or to both; and it may or may not be variable-binding. The general form of a formula whose outermost operator is Φ can thus be written as

$$\Phi\bar{x}.(\bar{A}, \bar{t}\,)$$

where $\bar{A}$ is a sequence of formulae, $\bar{t}$ is a sequence of terms, and $\bar{x}$ is a sequence of variables that are bound by the outermost (i.e., the displayed) Φ in this formula. Let us assume that Φ is associated with an introduction rule (so it is not introduced by explicit definition). A canonical derivation has the following form

$$\cfrac{\cfrac{\mathscr{D}_1 \quad \ldots \quad \mathscr{D}_n}{\Phi\bar{x}.(\bar{A}, \bar{t})}\ \Phi\text{-INTRO}}{\overline{\overline{\qquad B \qquad}}}\ \text{definitional substitutions}$$

The number of definitional substitutions may be zero, in which case we have a canonical derivation in the old sense, namely one that ends in the application of an introduction rule. If the number of definitional substitutions is greater than zero,

[5] Cf. the stronger result of Tait [19] that Gödel's T enjoys both of these properties.

then the conclusion B is in general not syntactically identical to, though it is definitionally identical to, $\Phi\bar{x}.(\bar{A},\bar{t})$. An important special case is where B has the form $\Phi\bar{x}'.(\bar{A}',\bar{t}')$.

It is quite immaterial precisely which definitional substitutions are applied in getting from $\Phi\bar{x}.(\bar{A},\bar{t})$ to B. In general, the only things we care about in a derivation Δ consisting entirely of definitional substitutions are the starting point (the leftmost formula) and the end point (the conclusion). We are therefore led to postulate all reductions of the following form:

$$\begin{array}{c}\mathscr{D}\\ A\\ \Delta\\ B\end{array} \quad\rightsquigarrow\quad \begin{array}{c}\mathscr{D}\\ A\\ \Delta'\\ B\end{array} \tag{Δ-red}$$

Here Δ and Δ' are derivations consisting entirely of definitional substitutions whose starting point is A and whose end point is B.

In getting from $\Phi\bar{x}.(\bar{A},\bar{t})$ to the definitionally identical $\Phi\bar{x}'.(\bar{A}',\bar{t}')$ it is natural to proceed by first substituting each of the items in the list $\bar{A},\bar{t}$ in the order given and thereafter rename the bound variables $\bar{x}$ to $\bar{x}'$. We shall call this the *regular* definitional substitution from $\Phi\bar{x}.(\bar{A},\bar{t})$ to $\Phi\bar{x}'.(\bar{A}',\bar{t}')$. As examples let us consider conjunction and universal quantification:

$$\dfrac{\dfrac{A\wedge B \qquad A\equiv A'}{A'\wedge B} \qquad B\equiv B'}{A'\wedge B'} \qquad\qquad \dfrac{\dfrac{\forall x A \qquad A\equiv A''}{\forall x A''}}{\forall x' A'}\,\alpha$$

In the right-hand derivation the only difference, if any, between A'' and A' is that A' has x' free wherever A'' has x free.

With a revised, or generalized, notion of canonical derivation, it is natural also to revise, or generalize, the standard reductions of first-order logic given by Prawitz [13]. Thus we wish to define reductions for derivations ending in the pattern:

introduction rule + definitional substitutions + elimination rule

Because of the reduction rule Δ-red, we may assume that we have a regular definitional substitution from the conclusion of the introduction rule to the major premiss of the elimination rule. The definition of the various reductions is then straightforward. For instance, for conjunction we have the following reduction:

$$\wedge\text{-ELIM}\;\dfrac{\dfrac{\dfrac{\wedge\text{-INTRO}\;\dfrac{\begin{array}{c}\mathscr{D}_1\\ A\end{array}\qquad\begin{array}{c}\mathscr{D}_2\\ B\end{array}}{A\wedge B}\qquad\begin{array}{c}\mathfrak{D}_1\\ A\equiv A'\end{array}}{A'\wedge B}\qquad\begin{array}{c}\mathfrak{D}_2\\ B\equiv B'\end{array}}{A'\wedge B'}}{A'} \quad\rightsquigarrow\quad \dfrac{\begin{array}{c}\mathscr{D}_1\\ A\end{array}\qquad\begin{array}{c}\mathfrak{D}_1\\ A\equiv A'\end{array}}{A'}$$

Note that, if $\mathscr{D}_1$ is canonical in the revised sense, then so is the derivation to the right here, since it is obtained by extending $\mathscr{D}_1$ with one definitional substitution. For the universal quantifier we have the following reduction:

$$\forall\text{-ELIM}\,\dfrac{\alpha\,\dfrac{\dfrac{\forall\text{-INTRO}\,\dfrac{\begin{array}{c}\mathscr{D}\\ A\end{array}}{\forall x A} \qquad \begin{array}{c}\mathfrak{D}\\ A \equiv A''\end{array}}{\forall x A''}}{\forall x' A'}}{A'[t/x']} \quad\rightsquigarrow\quad \dfrac{\begin{array}{c}\mathscr{D}[t/x]\\ A[t/x]\end{array} \qquad \dfrac{\begin{array}{c}\mathfrak{D}\\ A \equiv A''\end{array}}{A[t/x] \equiv A''[t/x]}\,\text{R2}}{A''[t/x]}$$

To see that this is indeed a reduction, notice that A'' differs from A' at most in having x free wherever A' has x' free; hence $A'[t/x']$ and $A''[t/x]$ are syntactically identical. Again it holds that if $\mathscr{D}$ is canonical in the revised sense, then so is the derivation on the right here.

Prawitz's definition of validity now applies just as before. In particular, a canonical derivation is said to be valid iff all its immediate sub-derivations are valid. It then follows that if $\mathscr{D}'$ is a valid canonical derivation, then so is the derivation $\mathscr{D}$:

$$\dfrac{\begin{array}{c}\mathscr{D}'\\ A\end{array} \qquad \begin{array}{c}\mathfrak{D}\\ a \equiv b\end{array}}{A[b/a]_!}$$

In other words, the extension of a valid canonical derivation by a definitional substitution is again a valid canonical derivation.

7 The Justification of =-ELIM

An application of =-INTRO followed by a regular definitional substitution has the following form:

$$\dfrac{\dfrac{t' = t' \qquad \begin{array}{c}\mathfrak{D}_1\\ t' \equiv t\end{array}}{t = t'} \qquad \begin{array}{c}\mathfrak{D}_2\\ t' \equiv u\end{array}}{t = u} \tag{=-CanDer}$$

This is also the form of a canonical derivation of $t = u$. Here t and u may be syntactically different terms, since in the theory of definitional identity we can derive formulae of the form $t \equiv u$, for syntactically different t and u (for instance, we may let $t \equiv u$ be an explicit definition). Our generalized conception of canonical derivation thus captures formulae of the form $t = u$.

We are now in a position to give an informal justification of =-ELIM. We proceed according to the following steps:

1* Assume that the major premiss of =-ELIM, viz. $t = u$, is the conclusion of a canonical derivation.
2* Use the resources provided by this canonical derivation together with the minor premiss $A[x, x]$ in order to justify $A[t, u]$.

The canonical derivation of $t = u$ may be assumed to have the form =-CanDer. From the two sub-derivations $\mathfrak{D}_1$ and $\mathfrak{D}_2$ in =-CanDer we get $t \equiv u$ by the symmetry and transitivity of definitional identity. From $A[x, x]$, moreover, we get, firstly,

$A[t, t]$ by instantiating t for the free variable x and, thereafter, $A[t, u]$ by definitional substitution. We may write the justification schematically as follows:

$$\dfrac{\text{instantiation}\ \dfrac{A[x,x]}{A[t,t]} \qquad \dfrac{t' \equiv t \qquad t' \equiv u}{t \equiv u}\ \text{symmetry, transitivity}}{A[t,u]}\ \text{definitional substitution}$$

In the final step it is of course essential that we rely on definitional substitution: had we relied instead on the indiscernibility of identicals, the justification would be circular.

For the more formal justification relying on Prawitz's notion of validity, we must define reduction for derivations of the form =-CanDer followed by an application of =-ELIM, just as we did for conjunction and the universal quantifier above.

$$\dfrac{\dfrac{\dfrac{t'=t' \qquad \overset{\mathfrak{D}_1}{t'\equiv t}}{t=t'} \qquad \overset{\mathfrak{D}_2}{t'\equiv u}}{t=u} \qquad \overset{\mathscr{D}[x]}{A[x,x]}}{A[t,u]} \quad \rightsquigarrow \quad \dfrac{\dfrac{\overset{\mathscr{D}[t']}{A[t',t']} \qquad \overset{\mathfrak{D}_1}{t'\equiv t}}{A[t,t']} \qquad \overset{\mathfrak{D}_2}{t'\equiv u}}{A[t,u]}$$

In the right hand derivation here we are applying definitional substitution, and not =-ELIM, in order to pass from $A[t', t']$ to $A[t, u]$. Therefore, if $\mathscr{D}[t']$ is a canonical derivation, then so is the whole derivation on the right.

Let us recall how the formal justification is to proceed. We assume that we have two valid derivations

$$\overset{\mathscr{D}_1}{t=u} \qquad \overset{\mathscr{D}_2}{A[x,x]}$$

On this assumption we must show that the derivation

$$\dfrac{\overset{\mathscr{D}_1}{t=u} \qquad \overset{\mathscr{D}_2}{A[x,x]}}{A[t,u]} \tag{=-ElimDer}$$

is valid. We may assume that =-ElimDer is a closed derivation. Our task is therefore to show that it reduces to a valid canonical derivation.

Since =-ElimDer is closed, it follows that $\mathscr{D}_1$ is closed; that there are no open assumptions in $\mathscr{D}_2$; and that the only variable free in $\mathscr{D}_2$ is x.

That $\mathscr{D}_1$ is valid therefore means that it reduces to a derivation of the form

$$\begin{array}{c} t'=t' \\ \Delta \\ t=u \end{array} \tag{$\mathscr{D}_1'$}$$

where Δ is a sequence of definitional substitutions. By the reduction rule Δ-red we may assume that Δ is regular, hence that $\mathscr{D}_1'$ is of the form =-CanDer.

On =-ElimDer we first reduce $\mathscr{D}_1$ to $\mathscr{D}_1'$ and thereafter use our revised =-reduction to obtain

$$\dfrac{\dfrac{\overset{\mathscr{D}_2[t']}{A[t',t']} \qquad \overset{\mathfrak{D}_1}{t'\equiv t}}{A[t,t']} \qquad \overset{\mathfrak{D}_2}{t'\equiv u}}{A[t,u]}$$

From the assumptions on $\mathscr{D}_2$ it follows that $\mathscr{D}_2[t']$ is closed and valid. It therefore reduces to a valid canonical derivation $\mathscr{D}_3$. Being canonical, this derivation ends in an introduction rule followed by some number of definitional substitutions; but then so does the derivation

$$\dfrac{\dfrac{\begin{matrix}\mathscr{D}_3\\ A[t',t']\end{matrix} \qquad \begin{matrix}\mathfrak{D}_1\\ t'\equiv t\end{matrix}}{A[t,t']} \qquad \begin{matrix}\mathfrak{D}_2\\ t'\equiv u\end{matrix}}{A[t,u]}$$

This is therefore a canonical derivation of $A[t,u]$; it is valid, since $\mathscr{D}_3$ is valid; and it is a derivation to which =-ElimDer reduces. Thus we have shown that =-ELIM is justified.

Acknowledgements The justification of =-ELIM given here is an adaptation to natural deduction of the justification of the corresponding rule in Martin-Löf's type theory given in [7]. Correspondence with Per Martin-Löf pertaining to the latter work suggested to me the conception of canonical derivation that lies at the basis of the current work. While writing the paper I have been financially supported by grant nr. 17-18344Y from the Czech Science Foundation, GAČR.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Curry, H.B. (1963). *Foundations of mathematical logic*. New York: McGraw-Hill.
2. Curry, H.B., & Feys, R. (1958). *Combinatory logic* Vol. 1. Amsterdam: North-Holland.
3. Fitch, F.B. (1952). *Symbolic logic*. New York: The Ronald Press.
4. Gentzen, G. (1933). Untersuchungen über das logische Schließen. *Mathematische Zeitschrift*, *39*, 176–210, 405–431.
5. Griffiths, O. (2014). Harmonious rules for identity. *Review of Symbolic Logic*, *7*, 499–510.
6. Kleene, S.C. (1952). *Introduction to metamathematics*. New York: Van Norstrand.
7. Klev, A. (2017). The justification of identity elimination in Martin-Löf's type theory. Topoi. Online First. https://doi.org/10.1007/s11245-017-9509-1.
8. Lemmon, E.J. (1965). *Beginning logic*. London: Nelson.
9. Martin-Löf, P. (1971). Hauptsatz for the intuitionistic theory of iterated inductive definitions. In Fenstad, J.E. (Ed.) *Proceedings of the second Scandinavian logic symposium* (pp. 179–216). Amsterdam: North-Holland.
10. Martin-Löf, P. (1975). About models for intuitionistic type theories and the notion of definitional equality. In Kanger, S. (Ed.) *Proceedings of the third Scandinavian logic symposium* (pp. 81–109). Amsterdam: North-Holland.
11. Martin-Löf, P. (1998). An intuitionistic theory of types. In Sambin, G., & Smith, J. (Eds.) *Twenty-five years of constructive type theory* (pp. 127–172). Oxford: Clarendon Press. First published as preprint in 1972.
12. Milne, P. (2007). Existence, freedom, identity, and the logic of abstractionist realism. *Mind*, *116*, 23–53.
13. Prawitz, D. (1965). *Natural deduction*. Stockholm: Almqvist & Wiksell.
14. Prawitz, D. (1971). Ideas and results in proof theory. In Fenstad, J.E. (Ed.) *Proceedings of the second Scandinavian logic symposium* (pp. 235–307). Amsterdam: North-Holland.

15. Prawitz, D. (1973). Towards a foundation of a general proof theory. In Suppes, P., Henkin, L., Joja, A., Moisil, G.C. (Eds.) *Logic, methodology and philosophy of science IV* (pp. 225–250). Amsterdam: North-Holland.
16. Read, S. (2004). Identity and harmony. *Analysis*, *64*, 113–119.
17. Read, S. (2016). Harmonic inferentialism and the logic of identity. *Review of Symbolic Logic*, *9*, 408–420.
18. Schroeder-Heister, P. (2006). Validity concepts in proof-theoretic semantics. *Synthese*, *148*, 525–571.
19. Tait, W.W. (1967). Intensional interpretations of functionals of finite type. *Journal of Symbolic Logic*, *32*, 198–212.