frontiers | Frontiers in Psychology

TYPE Conceptual Analysis
PUBLISHED 18 November 2022
DOI 10.3389/fpsyg.2022.1048975

OPEN ACCESS

EDITED BY
Peter beim Graben,
Humboldt University of Berlin,
Germany

REVIEWED BY
Hans Van Eyghen,
VU Amsterdam, Netherlands
Maria Mannone,
University of Palermo, Italy

*CORRESPONDENCE
Steven Phillips
steven.phillips@aist.go.jp

SPECIALTY SECTION
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

RECEIVED 20 September 2022
ACCEPTED 28 November 2022
PUBLISHED 18 November 2022

CITATION
Phillips S (2022) What is category theory to cognitive science? Compositional representation and comparison.
Front. Psychol. 13:1048975.
doi: 10.3389/fpsyg.2022.1048975

What is category theory to cognitive science? Compositional representation and comparison

Steven Phillips*

Mathematical Neuroscience Group, Human Informatics and Interaction Research Institute, National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

Category theorists and cognitive scientists study the structural (analogical) relations between domains of interest albeit in different contexts, that is, formal and psychological systems, respectively. Despite this basic commonality, very few cognitive scientists take a category theory approach toward understanding the structure of cognition which raises the question, What is category theory to cognitive science? An answer is given as the slogan "Category theory is to cognitive science as functor is to representation; as natural transformation is to comparison" to make category theory more accessible and informative for cognitive scientists.

KEYWORDS

category theory, category, functor, natural transformation, analogy, matrix reasoning

1. Introduction

What is category theory to cognitive science? A short answer is that both fields are about "comparison of (compositional) structure" albeit in different contexts. Category theory was invented to formalize correspondences between mathematical constructions (Eilenberg and Mac Lane, 1945; Mac Lane, 1998). Cognitive scientists often view cognition in terms of representations that preserve entity relationships (structure) *via* relationships between corresponding mental states: e.g., *classical compositionality* (Fodor and Pylyshyn, 1988). Despite contextual differences, category theory ideas relate to concepts in cognitive science in ways not generally recognized by cognitive scientists. Some basic connections between a mathematical theory of structure (i.e., category theory) and the structure of cognition are educed here for the purpose of making category theory more accessible and informative to cognitive scientists.

One might say that category theory and cognitive science share a common ideal: the representation and comparison of (compositionally) structured entities in some domain of interest. This situation is illustrated as the following square of arrows for the expression, *John loves Mary* (Figure 1). The relationship between the phrase and the concept is depicted as transporting or transforming the left vertical arrow, representing the structure of the expression, to the right vertical arrow, representing the conceptual structure, by sliding along the horizontal arrows, thus forming a square. This arrangement constitutes a so-called *commutative square* in that the chain of arrows in

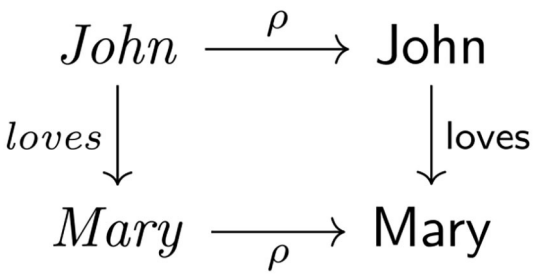

FIGURE 1
The expression "John loves Mary" as a commutative square of arrows.

the anticlockwise direction equals the chain of arrows in the clockwise direction, that is, $\rho \circ loves = \text{loves} \circ \rho$, where $\circ$ signifies the operation for combining arrows to form arrows. This arrangement is reminiscent of the *compositionality principle* (see, e.g., Janssen, 1997; Coecke et al., 2010) in linguistics linking syntax to semantics, or the *structure mapping theory* (Gentner, 1983) of analogy in cognitive psychology as a map from a source domain of knowledge to a target domain of knowledge, which features in a variety of analogy models (Gentner and Forbus, 2011) and a model of metaphor (Fuyama et al., 2020).

Despite the abstract nature of category theory, a substantial amount of the theory organizes around this notion of (typed) commutative square. In the context of cognition, basic category theory constructions correspond to compositional representation and comparison or transformation, which comports with the *representational/computational theory of mind* (Wilson, 1999) whereby cognition is seen as a system of computational processes over cognitive representations (see, Fodor, 1985, for a survey of this and other views). The purpose of this article is to make this consilience more concrete for cognitive scientists. The rest of this introduction is a preview to prime the details that follow.

Category theory vastly formalizes this simple idea of preserving structure as commutative squares with applications well beyond mathematics (see, e.g., Fong and Spivak, 2019). In recent years, researchers with a common interest in interdisciplinary applications of category theory have coalesced as the field known as *Applied Category Theory*. This "square" of arrows appears in many guises, historically, beginning with the formal concept of *natural transformation* (Eilenberg and Mac Lane, 1945), which depends on the concepts of *functor* and *category*. The relationships between these concepts are depicted as a diagram of arrows (Figure 2A). A category consists of *objects* and (directed) relations between objects, called *arrows*: e.g., f is an arrow from an object A to an object B, also written $f : A \to B$, in some category $\mathbf{C}$. The dotted lines indicate the actions of two functors on the objects and arrows in $\mathbf{C}$, that is, respectively, a functor F sends A, B, and f to objects $F(A)$ and $F(B)$ and arrow $F(f)$ in some category $\mathbf{D}$ and a functor G sends A, B, and f to the objects $G(A)$ and $G(B)$ and arrow $G(f)$ also in $\mathbf{D}$. Functors are arrows between objects that are categories: e.g., $F, G : \mathbf{C} \to \mathbf{D}$. The objects $F(A)$ and $F(B)$ and the arrow $F(f)$ constitute the *image* of F; likewise, the objects $G(A)$ and $G(B)$ and the arrow $G(f)$ constitute the image of G. The (naturality) square of arrows involving the images of F and G depicts a *natural transformation*, that is, a map η from F to G, written $\eta : F \dot{\to} G$, consisting of a component map $\eta_A : F(A) \to F(A)$ for each object A in $\mathbf{C}$ such that for each arrow f in $\mathbf{C}$ the square commutes. Thus, natural transformations are maps between functors and functors are maps between categories, hence the logical dependencies. Although originally introduced for applications in topology (Eilenberg and Mac Lane, 1945), category theory constructions also feature in diverse fields such as the use of natural transformations in music (Mannone and Favali, 2019) and categorical forms of compositionality in aesthetics (Kubota et al., 2017).

Categories, functors, and natural transformations partake in another important aspect of the general theory—*compositionality*: the composition of two *compatible* arrows (i.e., a pair of arrows linked "head to tail") is an arrow, for example, $f : A \to B$ composed with $g : B \to C$ is $g \circ f : A \to C$, where again $\circ$ signifies the composition operation. Composition operates at all levels: arrows between objects, functors between categories and natural transformations between functors. The diagram of arrows (Figure 2B) involves composition of arrows, $f : A \to B$ and $g : B \to C$, the action of functors on composed arrows, $F(g \circ f)$, and composition of natural transformations, $\alpha : F \dot{\to} G$ and $\beta : G \dot{\to} H$. (The diagram does not show functor composition, that is, in the third, out-of-plane direction.) These apparently different forms of compositionality are actually the same concept in different contexts (or dimensions): ordinary arrows are arrows between ordinary objects (vertical dimension), functors are arrows between objects that are categories, and natural transformations are arrows between objects that are functors in a category of functors (horizontal dimension). Natural transformations also compose with functors. Category theory provides a vast generalization of the notion of compositionality that is relevant to cognitive science.

Basic category theory concepts, though straightforwardly introduced this way, engender little intuition regarding applications as nothing is said about their specific nature. For cognitive scientists, however, these squares of relations between objects are also reminiscent of the squares of stimuli used in matrix reasoning tasks (Raven et al., 1998), which have been studied as tests of intelligence (Carpenter et al., 1990). Such reasoning tasks involve a matrix of stimuli with a missing cell that can be completed by applying the relationship educed from the rows or columns with all cells filled. For instance, the proportional analogy *Mare is to foal as cow is to what?* in matrix form involves a two-by-two matrix. The

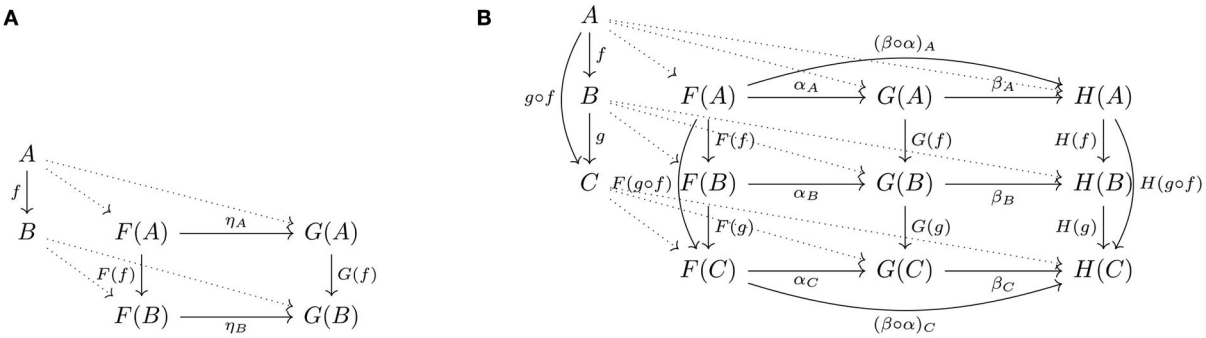

FIGURE 2
Natural transformations of **(A)** arrows and **(B)** compositions of arrows.

empty cell is filled with *calf* by educing that the relevant relationship between *mare* and *foal* is *gives-birth-to* and applying this relationship to *cow* to obtain *calf*. These and other such matrices of stimuli have a common form symbolized by the expression $a : b :: c : d$, where the semicolon corresponds to relationships in one (say, vertical) direction and the double semicolon to relationships in the other (horizontal) direction. This situation is analogous to a commutative square (Figure 3), which naturally extends to matrices with more rows and columns (cf. Figure 2B). Such comparisons afford a perspicuous way of bootstrapping intuitions about category theory concepts (Section 2) for potential applications in cognitive science (Section 3) to complement other conceptualizations (see, e.g., Phillips, 2021a, and the references therein).

Some readers may ponder the need for detailed explanations of basic category theory concepts given the many introductions that already exist. However, category theory introductions typically presume a style of thinking that can bedevil those outside the target audience seeking intuition (see Lawvere and Schanuel, 2009, for a general readership). For those readers, the theory can appear as a bridge to nowhere. Yet mathematics is arguably as much a reflection of thinking as it is about the world (Mac Lane, 1986; Lakoff and Núñez, 2000). The purpose here is to present category theory concepts to cognitive scientists in a way that enables thinking about thinking, categorically. Accordingly, this presentation departs from the usual style of relegating technical details to an appendix in favor of a side-by-side comparison to facilitate understanding—the devil is in the *comparable* details—which affords a novel synthesis of concepts for the purpose of doing cognitive science.

2. Some formal and conceptual comparisons

The formal basis of comparison employed here is the concept of natural transformation, which depends on the concept of functor, which in turn depends of the concept of category. Accordingly, categories are introduced first (Section 2.1), followed by functors (Section 2.2), and then natural transformations (Section 2.3).

2.1. Categories and compositional structure

Both category theory and cognitive science are concerned with modeling the *(compositional) structure* of some "domain" of interest, that is, the entities, the entity relationships, and the way that combinations for those relationships are themselves related. (The term domain is used in two senses: informally, as a topic of interest and, formally, as the source object of an arrow, Section 2.1.2). In cognitive science, for example, one may regard a cognitive process as consisting of a chain of subprocesses, a cognitive system as composed of subsystems, or a cognitive representation of some complex entity as constructed from representations of constituent entities. Category theory also considers a wide variety of compositional forms, including composition of functions, relations, algebraic structures, and spaces. All forms are based on a concept of compositionality, introduced here, that also pertains to situations of interest to cognitive scientists.

The concept of a category depends on several constituent concepts. Briefly, a category consists of a collection of *objects* and *arrows* (Section 2.1.1) where each arrow is directed from its *domain* object to its *codomain* object (Section 2.1.2), including self-directed *identity arrows* (Section 2.1.3), that combine as arrows by a *composition operation* satisfying certain properties (Section 2.1.4). A formal definition of category (Section 2.1.5) captures the notion of structure as an arrow: The collection of such structures is a category of such arrows (now objects), and the arrows between these objects are structure-preserving maps given as commutative squares (Section 2.1.6).

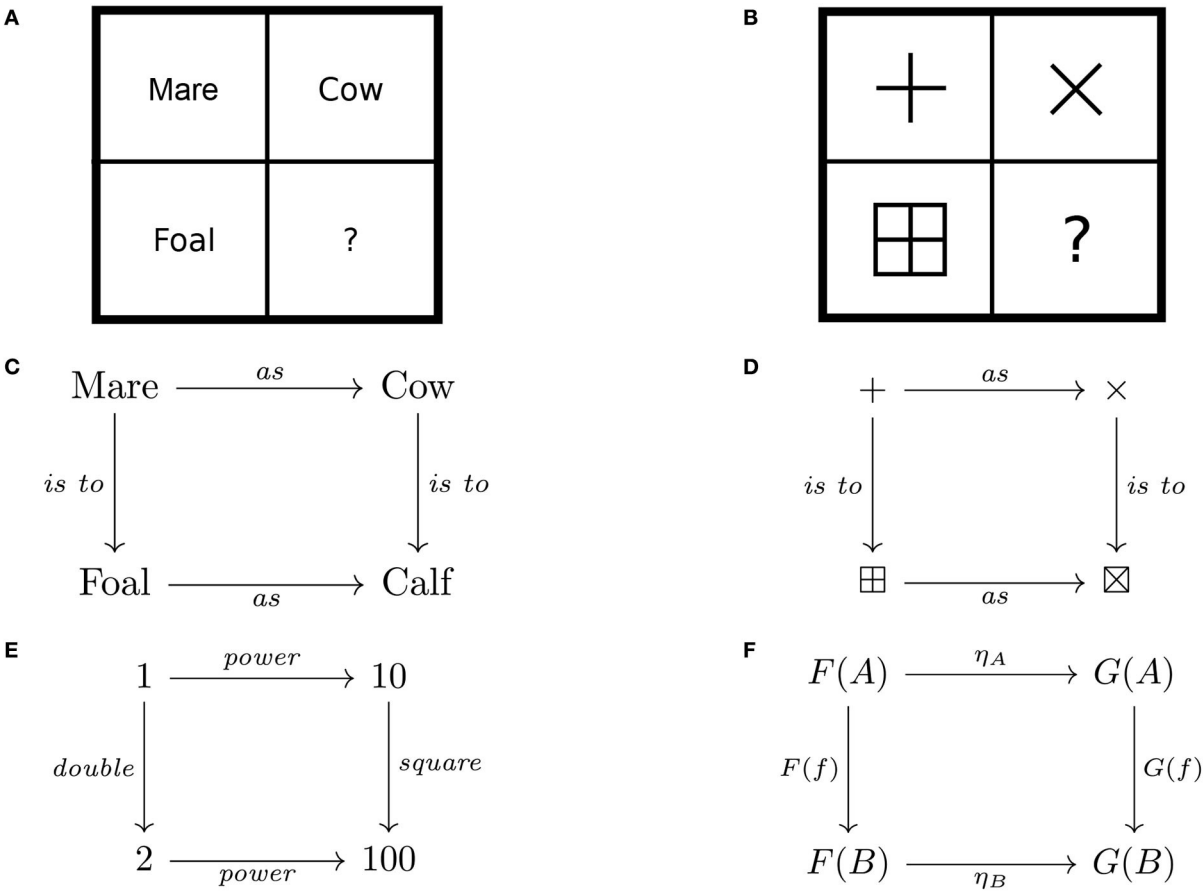

FIGURE 3
A comparison of proportional analogy, matrix reasoning, and natural transformation: **(A)** "Mare is to foal as cow is to calf" and **(B)** "+ is to ⊞ as × is to ⊠" in matrix form, and **(C,D)** in the form of a natural transformation. Algebraic relations, for example, **(E)** the power function are also instances of natural transformations, whose general form is given as **(F)** a commutative square of arrows.

2.1.1. Elements are related as objects and arrows

A domain of interest (such as a group of people) is typically considered in terms of its elements and relationships between those elements. An ordered set is a simple example of how one might model a domain this way. Suppose, for instance, three people of interest, *Ann*, *Bob*, and *Cal*, and that *Ann is shorter than Bob* and *Bob is shorter than Cal*. This domain can be modeled as an *ordered set*, that is, the set of people $P = \{Ann, Bob, Cal\}$ together with their order relationships $Ann < Bob$ and $Bob < Cal$. These elements and their order relationships correspond to instances of two basic kinds of constituents for a category, that is, each element corresponds to an instance of an abstract notion of entity, called *object*, and each order relationship to an instance of an abstract notion of relation between entities (objects), called *arrow* or *morphism* or *map*. So, for instance, the element *Ann* is now an object and the relationship $Ann < Bob$ is now an arrow $Ann \to Bob$ in some category corresponding to an ordered set.

Element is to relationship as object is to arrow.

The simplicity of this example belies several important subtleties with the move from sets to categories that are elaborated here and in subsequent sections. First, notice that the *less-than* symbol "points" to the shorter person whereas the arrow points to the taller person. Relationships are directed. Direction is captured syntactically (diagrammatically) by the direction of each arrow and semantically by two additional relations (maps) between objects and arrows, introduced in the next section (Section 2.1.2). What matters is that the relationship between syntax and semantics is consistent. Compare, for instance, the correspondence between order relationship $Ann > Bob$ and arrow $Ann \leftarrow Bob$.

Second, in general, order relationships are expressed using the "less-than *or* equals" symbol, $\leq$, rather than the less-than symbol. So, for instance, *Ann is shorter than or the same height as Bob*, now written $Ann \leq Bob$, means that *Ann is not taller than Bob*. The significance of this interpretation is that *Ann is not taller than Ann*, that is, the *self-directed* relationship $Ann \leq Ann$

which corresponds to the arrow $Ann \rightarrow Ann$. Order relations that are strictly less-than, that is, $a < b$ but not $a \leq b$ (whence, $a \not\leq a$), are called *strict orders*. For example, *parent-of* is a strict order. The relevance of this distinction to category theory will be elaborated shortly (Section 2.1.3).

Third, there may be more than one relationship between entities, in general, hence more than one arrow between a pair of objects. Notation referencing the arrow and the pair of objects serves to distinguish arrows in a category. Suppose, for instance, *Ann is younger than or the same age as Bob* is expressed by the order relationship $Ann \sqsubseteq Bob$, that is, *Ann is not older than Bob*. The arrows corresponding to the two order relationships between *Ann* and *Bob* are referenced in full as $\leq_{AB} : Ann \rightarrow Bob$ and $\sqsubseteq_{AB} : Ann \rightarrow Bob$, that is, different symbols or identifiers are used to distinguish different arrows between the same pair of object.

Fourth, and finally, category theory often affords more than one way to model a domain of interest. For instance, a (binary) relation R between sets A and B can be modeled as a collection of arrows between objects corresponding to the elements partaking in that relationship, as in the *Ann-Bill-Cal* example: there is an arrow $a \rightarrow b$ for each element $a \in A$ that is in an R relationship with an element $b \in B$. For this reason, such arrows are referred to as relationships, rather than relations. Or, as we will see later (Section 2.1.5, example 2), R can be modeled as a single arrow between those sets, $A \rightarrow B$. The term relation is used when a distinction between relation and relationship is not essential.

2.1.2. Relations are directed as domains before codomains

As mentioned earlier, the direction of each relationship is depicted by the direction of each arrow: e.g., $\leq_{AB} : Ann \rightarrow Bob$ is directed from object *Ann* to object *Bob*, which says that *Ann* comes before *Bob* and *Bob* comes after *Ann*. The objects *Ann* and *Bob* are called the *domain* and the *codomain* of the arrow $\leq_{AB}$, respectively.

Before is to after as domain is to codomain.

Expressions $\leq_{AB} : Ann \rightarrow Bob$ and $\leq_{AB} : Bob \leftarrow Ann$ identify the same arrow, whose directionality is determined by two maps between objects and arrows. In general, for a category $\mathbf{C}$, the collection of $\mathbf{C}$-objects is denoted $\mathbf{C}_0$ and the collection of $\mathbf{C}$-arrows is denoted $\mathbf{C}_1$. Two maps from arrows to objects $dom : \mathbf{C}_1 \rightarrow \mathbf{C}_0$ and $cod : \mathbf{C}_1 \rightarrow \mathbf{C}_0$ determine the domain and codomain object of each arrow, respectively. In the *Ann-Bob-Cal* example, the corresponding category, denoted *ABC*, consists of the set of objects $ABC_0 = \{Ann, Bob, Cal\}$, the set of arrows $ABC_1 = \{\leq_{AB}, \leq_{BC}, \dots\}$ and mappings that include $dom : \leq_{AB} \mapsto Ann$ and $cod : \leq_{AB} \mapsto Bob$.

Notice that nothing is said about the nature of object and arrow beyond their relationship to each other. How one is supposed to interpret these formal concepts depends on context, that is, the category in which they reside. For instance, in the context of sets and functions, an object is a set and an arrow is a function. Category theory can been seen as an abstraction of set theoretical constructions; hence, the notation and nomenclature are often taken from there: e.g., arrows (morphisms, maps) are generally written $f : A \rightarrow B$ even though objects need not be sets and arrows need not be functions (homomorphisms, mapping elements in a way that preserves their relationships).

2.1.3. Self-directed relations as identity arrows

As also mentioned earlier, order relationships are typically expressed using the $\leq$ symbol. Thus, $A \leq B$ says that A comes no later than B. For the *Ann-Bob-Cal* example, this situation means that each person is ordered with respect to themselves: e.g., $Ann \leq Ann$ means that *Ann* is not taller than herself. Hence, the set of arrows for the corresponding category, *ABC*, includes the self-directed arrows $Ann \rightarrow Ann$, $Bob \rightarrow Bob$ and $Cal \rightarrow Cal$. A relation R on A is called *reflexive* (has the reflexivity property) if every element $a \in A$ is related to itself. Reflexivity of order corresponds to an arrow $\leq_A : A \rightarrow A$ for each object A in some ordered set as a category. The arrow $\leq_A : A \rightarrow A$ is called the *identity arrow* at A.

Relationship is to reflexivity as arrow is to identity.

Every object A in a category is associated with an identity arrow, written $1_A : A \rightarrow A$. For example, in the context of sets and functions, the identity arrow at set A is the identity function $1_A : a \mapsto a$. The reason for denoting the identity arrow as 1_A is by analogy to multiplication of a number by 1 (Section 2.1.4). A self-directed arrow need not be an identity arrow. For instance, the constant function on a set A sending every element $a \in A$ to the same element $k \in A$, that is, $f : A \rightarrow A; a \mapsto k$, is self-directed but not an identity function. The composition of two compatible arrows is an arrow (Section 2.1.4). The rules for composition imply that every object in a category is associated with one and only one identity arrow. Thus, for a category $\mathbf{C}$, there is another relation between objects and arrows that is given by the map $id : \mathbf{C}_0 \rightarrow \mathbf{C}_1; A \mapsto 1_A$.

The implication that every object is associated with one identity arrow may seem too restrictive for cognitive science in situations where the entities do not have self-directed relationships and thus lack a meaningful interpretation in terms of the identity arrows of a category. However, as explained later (Section 2.1.6), these situations can be modeled by other categories.

2.1.4. Transitivity of relations as composition of arrows

Order relationships are themselves related to each other in a way that is called *transitivity*. For instance, *Ann is shorter*

than Bob and *Bob is shorter than Cal* implies *Ann is shorter than Cal*. All triples of ordered elements are related this way, that is, the transitivity property of order relations. Formally, a relation R between two sets A and B is a subset of the set of all pairs of elements with the first element of each pair drawn from A and the second element of each pair drawn from B, called the *Cartesian product* of A and B, that is, the set $R \subseteq A \times B = \{(a,b)|a \in A, b \in B\}$. If the pair (a,b) is in R, then we say that a is *R-related* to b, or write aRb to indicated this relationship. A relation R on A, that is, a subset of $A \times A$, is called *transitive* (has the transitivity property) if aRa' and $a'Ra''$ implies aRa'' for all triples of elements a, a', a'' in A. Transitivity of order corresponds to conjunction of arrows: If there is an arrow $A \to B$ *and* an arrow $B \to C$, then there is a *composite* arrow $A \to C$. This conjunction of arrows is a form of *composition*.

Relationship (order) is to transitivity as arrow is to composition (conjunction).

In any category **C**, if $f : A \to B$ and $g : B \to C$ are a pair of *compatible* arrows, that is, the codomain of the first arrow, f, is the domain of the second arrow, g, then there is a composite arrow from A to C in **C**, written $g \circ f : A \to C$, where $\circ$ denotes the *composition operation*, simply called *composition*. Composition is a (partially defined) map sending each pair of compatible arrows to their composition, that is, the map $comp : \mathbf{C}_1 \times \mathbf{C}_1 \to \mathbf{C}_1; (f,g) \mapsto g \circ f$, also denoted $\circ(-,-)$, cf. expressions $+(1,2)$ and $1+2$. For sets and functions, composition of compatible functions $f : A \to B$ and $g : B \to C$ is the composite function $g \circ f : A \to C; a \mapsto g(f(a))$ mapping each element $a \in A$ to the element $g(f(a)) \in C$, read *g of f of a*, hence the notational order. Note that in this context (category) of sets and functions, composition is a function *on* functions sending f and g to some function h in that category, that is, $g \circ f = h$, cf. $1 + 2 = 3$. The corresponding form for composition of order arrows $\leq_{AB} : A \to B$ and $\leq_{BC} : B \to C$ is $\leq_{AC} = \leq_{BC} \circ \leq_{AB} : A \to C$, where composition takes on the role of conjunction. In this context (ordered set as a category), composition is a function *on* order relationships. This situation compares with *transitive inference* as the logical rule of replacement $aRb \wedge bRc \Rightarrow aRc$, where $\wedge$ is conjunction.

Composition of orders and functions satisfy two important properties that are required of a composition operation on the collection of arrows in any category, generally: *associativity* (Section 2.1.4) and *unity* (section 2.1.4).

2.1.4.1. Composition is associative: Composition order is commutative

Suppose the *Ann-Bob-Cal* example is extended to include a fourth person, *Dan*, and the order relationship *Cal is shorter than Dan*. Transitive inference can be applied twice in two ways to infer than *Ann is shorter than Dan*: (1a) $Ann \leq Bob$ and $Bob \leq Cal$ implies $Ann \leq Cal$ and (1b) $Ann \leq Cal$ and $Cal \leq Dan$ implies $Ann \leq Dan$, or (2a) $Bob \leq Cal$ and $Cal \leq Dan$ implies $Bob \leq Dan$ and (2b) $Ann \leq Bob$ and $Bob \leq Dan$ implies $Ann \leq Dan$. Compare this logical equivalence with the equality of the corresponding arrows, that is, a comparison of

- $aRb \wedge (bRc \wedge cRd) \Leftrightarrow (aRb \wedge bRc) \wedge cRd$ and
- $\leq_{CD} \circ (\leq_{BC} \circ \leq_{AB}) = (\leq_{CD} \circ \leq_{BC}) \circ \leq_{AB}$.

The order of compositions does not affect the result, that is, conjunction is *associative*. Associativity is also a property of addition, that is, $x + (y + z) = (x + y) + z$.

Associativity is to addition as associativity is to composition.

In any category **C**, the composition operation is associative, that is, $h \circ (g \circ f) = (h \circ g) \circ f$ for all triples of compatible arrows f, g, h in **C**. Brackets can be omitted, since the result is not affected by order of composition—cf. $h \circ g \circ f$ and $x + y + z$.

Associativity of composition is essentially commutativity of composition order (Figure 4). The "commutativity" of commutative squares is analogous to the commutativity of addition given as a square of arrows corresponding to numbers (Figure 4A). Addition can also be expressed as a square of operations between numbers (Figure 4B) or set of numbers (Figure 4C). A commutative composition operation is analogous to the commutativity of addition (Figure 4D), but composition is generally not commutative. However, associativity of composition can be expressed as a commutative square (Figure 4E), which in turn is expressed as a commutative square of operations on (hom-)sets of arrows (Figure 4F). Thus, the order of composition is commutative, hence the analogy between associativity and commutativity.

Commutativity is to addition as associativity is to order of composition.

2.1.4.2. Composition with identity arrows is unital

Identity arrows, introduced earlier (Section 2.1.3), play a special role with respect to composition just like the number 1 plays a special role with respect to multiplication.

One is to multiplication as identity is to composition.

Self-directed order arrows, $\leq_A$, also play an analogous role. Compare

- $x \times 1 = x = 1 \times x$,
- $\leq_{AB} \circ \leq_A = \leq_{AB} = \leq_B \circ \leq_{AB}$ and
- $f \circ 1_A = f = 1_B \circ f$.

For any category **C**, the composition operation has this property, called *unity*, that is, $f \circ 1_A = f = 1_B \circ f$ for all arrows f in **C**.

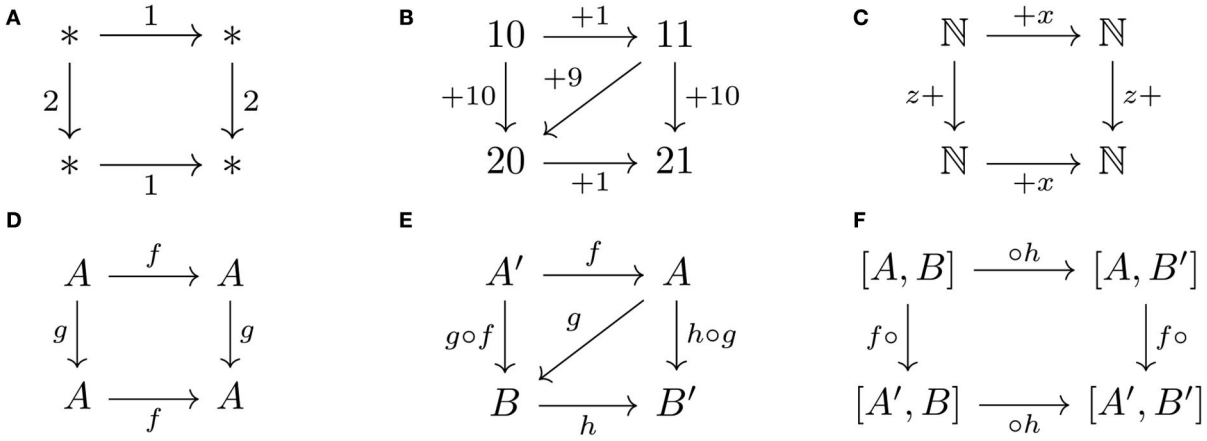

FIGURE 4
Commutativity and associativity properties are expressed as commutative squares: **(A)** commutativity of addition, **(B,C)** associativity of addition, **(D)** commutative composition operation, and **(E,F)** associativity of composition operation.

2.1.5. Formal constructions formalized as categories

The previous sections introduced all the basic concepts that make up a formal definition of category (Definition 1), that is, the concepts of object, arrow, domain, codomain, identity arrow, composition operation and the associativity, and unity properties for composition. A formal definition of category is introduced next followed by examples that further exercise this concept.

Definition 1 (category). *A category* $\boldsymbol{C} = (\boldsymbol{C}_0, \boldsymbol{C}_1, dom, cod, id, \circ)$ *consists of*

- *a collection of objects,* $\boldsymbol{C}_0 = \{A, B, C, \dots\}$,
- *a collection of arrows,* $\boldsymbol{C}_1 = \{f, g, h, \dots\}$—*an arrow* f *is directed from an object* A *to an object* B, *written* $f : A \to B$, *called the domain and codomain of* f, *respectively,*
- *two maps* $dom, cod : \boldsymbol{C}_1 \to \boldsymbol{C}_0$ *sending each arrow* $f : A \to B$ *to its domain and codomain object, respectively, that is,* $dom(f) = A$ *and* $cod(f) = B$,
- *a map* $id : \boldsymbol{C}_0 \to \boldsymbol{C}_1$ *assigning to each object* A *an arrow* $1_A : A \to A$, *called the identity arrow at* A, *that is,* $id(A) = 1_A$, *and*
- *a composition operation,* $\circ$, *sending each pair of compatible arrows* $f : A \to B$ *and* $g : B \to C$, *that is,* $cod(f) = dom(g)$, *to the arrow* $g \circ f : A \to C$ *that is*
  - *associative:* $h \circ (g \circ f) = (h \circ g) \circ f$, *and*
  - *unital:* $f \circ 1_A = f = 1_B \circ f$

for all compatible arrows $f, g, h \in \boldsymbol{C}_1$. *The collection of arrows in* $\boldsymbol{C}$ *with domain* A *and codomain* B *is called a hom-set, written* $\mathrm{Hom}_{\boldsymbol{C}}(A, B)$, $\boldsymbol{C}(A, B)$, *or* $[A, B]$ *when the category is clear from context.*

Several examples of categories were already given in the previous section to illustrate the basic concepts. These and closely related examples are listed for comparison (Example 2).

Example 2 (sets, functions, relations). *The following are categories.*

1. *The category* ***Set*** *has sets for objects and (total) functions between sets for arrows. The identity arrows are the identity functions. Composition is composition of functions.*
2. *Restricting the collection of functions to be inclusions,* $A \subseteq B$, *yields another category, denoted* $\boldsymbol{Set}^{\subseteq}$. *The identity arrows are* $A \subseteq A$.
3. *Every set* S *corresponds to a category whose objects are the elements* $s \in S$ *and only arrows are the identity arrows* $1_s : s \to s$. *Categories with only identity arrows are called discrete categories. Composition is trivial:* $1_s \circ 1_s = 1_s$. *Certain sets play an important role, such as the empty set,* $\emptyset$, *singleton set, written* $\{*\}$ *when the identity of the only element is not needed, and the set of natural numbers,* $\mathbb{N}$. *These sets correspond to important categories.*
   (a) *The empty category, denoted* ***0*** *(or 0), has no objects or arrows. Composition is the empty map, that is,* $\emptyset \times \emptyset \to \emptyset$. *Likewise, dom, cod and id are empty maps.*
   (b) *The singleton category, denoted* ***1*** *(or 1), has one object and one (identity) arrow.*
   (c) *An index category, denoted* $I \subseteq \mathbb{N}$, *has a subset of the natural numbers as objects.*
4. *An ordered set* $(P, \leq)$ *corresponds to a category whose objects are the elements* $p \in P$ *with an arrow* $p \to q$ *whenever* $p \leq q$.

Identity arrows correspond to reflexivity and composition to transitivity.

5. *The category* ***Rel*** *has sets for objects and relations between sets for arrows. The identity arrows are the identity relations, that is,* $1_A = \{(a,a)|a \in A\}$. *The composition of arrows is defined by the join of the corresponding relations, that is, the join of relations* $R \subseteq A \times B$ *and* $S \subseteq B \times C$ *is the relation* $R \bowtie S = \{(\acute{\pi}(r), \grave{\pi}(s))|r \in R, s \in S, \grave{\pi}(r) = \acute{\pi}(s)\}$, *where* $\acute{\pi}(a,b) = a$ *and* $\grave{\pi}(b',c) = c$, *that is,* $\acute{\pi}$ *and* $\grave{\pi}$ *return the left (first) and right (second) elements of each pair, respectively. In other words, a pair* (a,c) *is in the join of R and S whenever there exists a pair* (a,b) *in R and a pair* (b',c) *in S such that* $b = b'$.

Functions and relations are closely connected. Every function corresponds to a relation, and every relation corresponds to a set-valued function (Remark 3). Relations can be used to model non-determinism: Each element is sent to a set of possible outcomes.

Remark 3. *The graph of a function* $f : A \rightarrow B$ *is the relation* $\Gamma(f) = \{(a, f(a))|a \in A)\} \subseteq A \times B$. *A relation* $R \subseteq A \times B$ *corresponds to the set-valued (partial) function* $f_R : a \mapsto \{b|(a,b) \in R\}$ *defined on the subset of R-related elements of A.*

The join of two relations is analogous to compatibility of arrows. Accordingly, composition can be defined as a total function on the pairs of compatible arrows (Remark 4).

Remark 4. *For a category* $\boldsymbol{C}$, *the collection of pairs of compatible arrows is given by a constrained product on the collection of arrows, that is,* $\boldsymbol{C}_1 \times_{\boldsymbol{C}_0} \boldsymbol{C}_1 = \{(f,g)|cod(f) = dom(g), f \in \boldsymbol{C}_1, g \in \boldsymbol{C}_1\}$. *In this case, composition is the (total) map* $comp : \boldsymbol{C}_1 \times_{\boldsymbol{C}_0} \boldsymbol{C}_1 \rightarrow \boldsymbol{C}_1$.

Another alternative defines composition as a family of (total) maps indexed by triples of objects in the category (remark 5).

Remark 5. *For a category* $\boldsymbol{C}$, *the composition of arrows can be defined for each triple of objects* A, B, C *in* $\boldsymbol{C}$ *as the total function* $\circ_{ABC} : [A,B] \times [B,C] \rightarrow [A,C]; (f,g) \mapsto g \circ f$.

Notice that if the triple of objects is (A,A,A), then composition is just composition of self-directed arrows, that is, $\circ_{AAA} : [A,A] \times [A,A] \rightarrow [A,A]; (a,a') \mapsto a' \circ a$. And since composition is associative and unital, we also have $a \circ (a' \circ a'') = (a \circ a') \circ a''$ and $a \circ 1 = a = 1 \circ a$ where 1 is the identity arrow at A. As we have already seen, this situation is analogous to multiplication for numbers: e.g., $\times_{\mathbb{N}} : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}; (x,y) \mapsto x \cdot y$. Multiplication is also associative and unital: $x \cdot (y \cdot z) = (x \cdot y) \cdot z$ and $x \cdot 1 = x = 1 \cdot x$. This situation corresponds to an important algebraic structure, called a *monoid*. Restricting the category to just a single object introduces another example of category, where the self-directed arrows play the role of the elements, such as numbers, and composition plays the role of the binary operation, such as multiplication (Example 6).

Example 6 (monoid). *A monoid* $(M, \cdot, e)$ *consists of a set M, a binary operation* $\cdot$, *and an element* $e \in M$ *satisfying*

- *associativity:* $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ *and*
- *unity:* $a \cdot e = a = e \cdot a$

for all elements $a, b, c \in M$. *For instance, the set of real numbers together with elementary multiplication and 1 constitute a monoid,* $(\mathbb{R}, \times, 1)$. *The binary operation is called multiplication, and e is called the unit. These concepts are abstract notions. Any set and binary operation satisfying these axioms is a monoid, which includes the real numbers together with elementary addition as the "multiplication" operation and 0 as the "unit" element,* $(\mathbb{R}, +, 0)$. *Every monoid corresponds to a one-object category whose elements* $a \in M$ *correspond to the arrows* $a : * \rightarrow *$ *and binary operation to the composition operation, that is,* $a \cdot b$ *corresponds to the composite* $b \circ a : * \rightarrow *$. *The unit, e, corresponds to the identity arrow.*

2.1.6. Categories of structures as objects and structural relations as arrows

The examples introduced so far involved categories of objects that were elements or sets of elements with no other "internal" structure, that is, there were no relationships between the constituents of the objects themselves. However, an entity may have additional internal structure that requires modeling. For instance, similarity judgments may be regarded as a map on a set of related perceptual states so that similar stimuli evoke similar responses. One approach to this situation is to model the internal structure of an entity in the same way as domains are modeled, that is, as categories that are now objects in a "larger" category. Yet, this approach can be too restrictive as a collection of objects that are not categories individually may still constitute a category collectively. An alternative approach is to consider a weaker notion of structure, where the relations between constituents need not be reflexive (identity arrows are not required) or transitive (composition of arrows is not required). For instance, *parent-of* is neither reflexive, nor transitive. An efficient way of constructing categories for such situations is to use the internal structure of an existing category, which can circumvent the need to prove that the new construction satisfies the requirements of a category. A prototypical example is to construct the category of graphs and graph homomorphisms from the category of sets and functions, **Set** (Section 2.1.6). Constructing categories from a progenitor category, such as **Set**, can also be employed to model structural relations between domains of interest: If the constructed object is also a category, then maps between such objects correspond to maps between categories. A prototypical example of this situation is the category of monoids and monoids homomorphisms (Section 2.1.6). When the target object is a general category, this approach leads to a definition

of map between category, called *functor*, which is taken up later (Section 2.1.6).

2.1.6.1. Internal structure

A weaker notion of structure is *directed graph*, simply called *graph* hereafter, which consists of a set of *vertices*, a set of directed *edges* between vertices, and two maps determining the *source* and *target* vertex of each edge (Definition 7). Since a graph is given by sets and functions, graphs can be constructed from the internal structure of **Set**. A graph need not have a *loop* for each vertex (i.e., an edge whose source and target are the same vertex), or an edge for each (connected) *path*. A relation R on a set A corresponds to a graph with a vertex for each element $a \in A$ and an edge from a to a' for each relationship aRa'. Hence, graphs correspond to relations that need not be reflexive or transitive.

Definition 7 (graph). *A (directed) graph* $G = (G_0, G_1, src, tgt)$ *consists of*

- *a set of vertices,* $G_0 = \{v, w, \dots\}$,
- *a set of edges,* $G_1 = \{e, f, \dots\}$—*an edge* e *is directed from a vertex* v *to a vertex* w, *written* $e : v \to w$, *called the source and target of* e, *respectively, and*
- *two maps,* $src, tgt : G_1 \to G_0$ *sending each edge* $e : v \to w$ *to its source and target vertex, respectively, that is,* $src(e) = v$ *and* $tgt(e) = w$.

A comparison of the definitions for graph and category makes clear that graph is indeed a weaker notion of structure than category—every category corresponds to a graph by regarding the objects as vertices and the arrows as edges; however, not every graph can be interpreted as a category, because a graph may not have a loop for each vertex corresponding to an identity arrow for each object, or an edge for each path corresponding to an arrow for each composition of arrows.

Perhaps less clear is whether a collection of graphs (as objects) constitutes a category. To clarify this point, we need to specify the relations between graphs (as arrows), how those relations compose (as the composition operation), and whether composition satisfies associativity and unity. (The other relations are usually easy to provide, that is, *dom*, *cod* and *id* maps.) The archetypal relation between graphs is *graph homomorphism*, that is, a pair of maps $(h_0, h_1) : G \to G'$ sending each vertex v in G to the vertex $h_0(v)$ in G' and each edge $e : v \to w$ in G to the edge $h_1(e) : h_0(v) \to h_0(w)$ in G' (see Remark 8). The identity arrows are pairs of identity maps on the sets of vertices and edges. To show that this arrangement constitutes a category is straightforward, albeit tedious, as one needs to show composition produces a graph homomorphism satisfying associativity and unity.

A more efficient approach is to construct the graph from another category, which carries over the needed properties of associativity and unity, that is, to use the internal structure of another category. Observe that a graph is given by a pair of sets and a pair of functions. Thus, graphs are constructed from the internal structure of the category of sets and functions, **Set** (Remark 8).

Remark 8. *A graph* $G = (G_0, G_1, src, tgt)$ *can be expressed as a pair of arrows (Figure 5A). Hence, a graph homomorphism* $h : G \to G'$ *is expressed by the pair of commutative squares (Figure 5B), that is, a pair of maps* (h_0, h_1) *such that* $h_0 \circ src = src' \circ h_1$ *and* $h_0 \circ tgt = tgt' \circ h_1$.

This approach can be generalized further by making use of the analogy between vertices/edges and objects/arrows, that is, an arrow corresponds to a graph with domain and codomain objects corresponding to source and target vertices. We just saw how a graph is an object and an arrow between graphs (as objects) is a graph homomorphism. Accordingly, an arrow is now an object and an arrow between such objects is now an arrow homormorphism given by a commutative square of arrows in some category (example 9).

Example 9 (arrows). *The category* ***Arr(C)*** *has for objects the arrows* $\alpha : A_s \to A_t$ *of* $\mathbf{C}$ *and for arrows the arrow homomorphisms* $f : \alpha \to \beta$, *that is, pairs of arrows* (f_s, f_t) *of* $\mathbf{C}$ *such that the corresponding diagram commutes (Figure 5C). The identities* 1_α *in* ***Arr(C)*** *are the pairs of identity arrows* $(1_{A_s}, 1_{A_t})$. *Composition in* ***Arr(C)*** *is composition of commutative squares, which follows from the associativity and unity properties of the composition operation,* $\circ$, *in the original category* $\mathbf{C}$.

This approach also applies to ordered sets. Recall (Section 2.1.3) that a strictly ordered set is not a category because it lacks reflexivity, hence the corresponding identity arrows. However, a collection of strictly ordered sets is a category in the same way as for graphs. A strictly ordered set is a special case of a graph that lacks loops or cycles, that is, edges or paths from/to the same vertex v corresponding to the relationship $v \leq v$, and at most one edge from each source to each target. The objects are strictly ordered sets and the arrows are *monotonic functions*, that is, a function $f : P \to Q$ such that $p < p'$ implies $f(p) < f(p')$ for all elements $p \in P$. For comparison, a monotonic function on a non-strict ordered set, $(P, \leq)$, satifies $p \leq p'$ implies $f(p) \leq f(p')$ for all $p \in P$.

2.1.6.2. External structure

If the object of interest is already a category, then the maps between these objects constitute a notion of "external" category structure, that is, maps between categories, taken up in the next section (Section 2.2). For example, a monoid can also be expressed in terms of sets and functions (Example 10).

Example 10 (monoids). *Suppose monoids* $(M, \cdot_M, e_M)$ *and* $(N, \cdot_N, e_N)$. *A monoid homomorphism is a function* $h : M \to N$ *that preserves the*

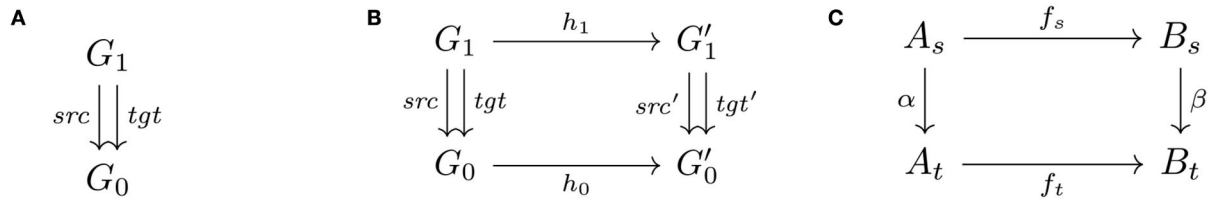

FIGURE 5
(A) Graphs, **(B)** graph homomorphisms, and **(C)** arrow homomorphisms are expressed as arrows.

- *binary operation:* $h(a \cdot_M b) = h(a) \cdot_N h(b)$ *and*
- *unit:* $h(e) = e_N$.

A monoid $(M, \cdot, e)$ *is equivalently given as the triple* (M, μ, η)*, where* $\mu : M \times M \to M$ *is the binary operation expressed as a bivariate function, that is,* $\mu : (a, b) \mapsto a \cdot b$*, and* $\eta : 1 \to M$ *is the unit expressed as a nullary function, that is,* $\eta : * \mapsto e$*. Thus, a monoid corresponds to a sum of arrows* $\mu + \eta : M \times M + 1 \to M$*, where the addition symbol signifies disjoint union of sets and functions, that is,* $A + B = \{(1, a) | a \in A\} \cup \{(2, b) | b \in B\}$ *with the corresponding sums of functions (Figure 6A), and a monoid homomorphism* $h : M \to N$ *corresponds to a commutative square (Figure 6B). The associativity and unity conditions for a monoid are also expressed as commutative diagrams.*

2.2. Functors and representation

Like compositionality, some notion of *representation* is central to category theory and cognitive science. A cognitive representation is usually taken to mean a mental state that stands in some correspondence relation to a state of the world, which can include other mental states, and a compositional (cognitive) representation means that the relationships between constituent mental states correspond to relationships between the constituent states of the world. Suppose, for example, a state of the world that has *John is to the left of Mary*. Viewing cognition as a *language of thought* (Fodor, 1975), for instance, supposes at least a symbol for *John*, John, and a symbol for *Mary*, Mary, that are juxtaposed in such a way that the spatial relationship between *John* and *Mary* is expressed by the syntactic relationship between their corresponding symbols: e.g., the pair of symbols (John, Mary). These symbolic representations are supposed to map to corresponding brain states by a *physical instantiation mapping* that likewise preserves the corresponding relations (see, e.g., Fodor and Pylyshyn, 1988, footnote 9). Category was introduced as a form of compositionality. A map between categories is called a *functor* preserving categorical structure. Functors afford a category theory notion of compositional cognitive representation and instantiation.

The value of casting the definition of category as collections of objects and arrows and their structural relations in terms of maps between those collections now becomes apparent. A map between categories, that is, a functor, is straightforwardly just a homomorphism preserving those structural relations, specified by equality conditions (Definition 11).

Definition 11 (functor). *Suppose categories* $\boldsymbol{C}$ *and* $\boldsymbol{D}$*. A functor* $F : \boldsymbol{C} \to \boldsymbol{D}$ *is a pair of maps* $(F_0, F_1) : (\boldsymbol{C}_0, \boldsymbol{C}_1) \to (\boldsymbol{D}_0, \boldsymbol{D}_1)$ *that preserves*

- *domains and codomains:* $dom(F_1(f)) = F_0(dom(f))$ *and* $cod(F_1(f)) = F_0(cod(f))$*,*
- *identities:* $F_1(1_A) = 1_{F_1(A)}$ *and*
- *compositions:* $F_1(g \circ f) = F_1(g) \circ F_1(f)$

for all objects $A \in \boldsymbol{C}_0$*, arrows* $f \in \boldsymbol{C}_1$ *and pairs of compatible arrows* $(f, g) \in \boldsymbol{C}_1 \times \boldsymbol{C}_1$.

The equivalent commutative diagrams for the equality conditions (in definition 11) make plain that a functor is simply a (homo)morphism of the maps that constitute the structure of a category (remark 12).

Remark 12. *The conditions for a functor are equivalently given by commutative diagrams. The conditions for domains and codomains are given by commutative squares, where* $dom1 = id \circ dom$ *and* $cod1 = id \circ cod$ *(Figure 7A). The conditions for identity (Figure 7B), and composition (Figure 7C) are likewise expressed this way. So, a category is given as the sum of four maps that specify: (1) the composition operation,* $comp : \boldsymbol{C}_1 \times_{\boldsymbol{C}_0} \boldsymbol{C}_1 \to \boldsymbol{C}_1$*, (2) the domain of each arrow in the category as the associated identity arrow,* $dom1 : \boldsymbol{C}_1 \to \boldsymbol{C}_1$*, (3) the codomain of each arrow in the category as the associated identity arrow,* $cod1 : \boldsymbol{C}_1 \to \boldsymbol{C}_1$*, and (4) the identity arrow associated with each object in the category,* $id : \boldsymbol{C}_0 \to \boldsymbol{C}_1$*. The sum of these four arrows is the arrow* $struct = comp + dom1 + cod1 + id$*, where the addition symbol expresses alternative maps (see Example 14), that is,* $struct : \boldsymbol{C}_\bullet \to \boldsymbol{C}_1$*, where* $\boldsymbol{C}_\bullet$ *denotes* $\boldsymbol{C}_1 \times_{\boldsymbol{C}_0} \boldsymbol{C}_1 + \boldsymbol{C}_1 + \boldsymbol{C}_1 + \boldsymbol{C}_0$*. (*$\boldsymbol{C}_1 + \boldsymbol{C}_1$ *is also written* $2\boldsymbol{C}_1$*.) Categories* $\boldsymbol{C}$ *and* $\boldsymbol{D}$ *are given by arrows* $struct_{\boldsymbol{C}}$ *and* $struct_{\boldsymbol{D}}$*, respectively, and a map from* $struct_{\boldsymbol{C}}$ *to* $struct_{\boldsymbol{D}}$ *is a commutative square (Figure 7D). Hence, a functor is a category homomorphism.*

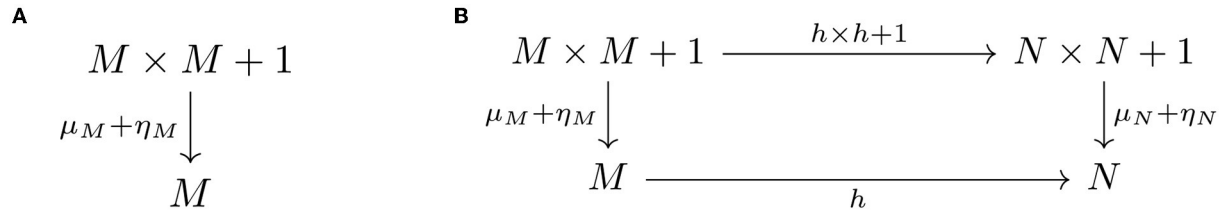

FIGURE 6
(A) Monoid and **(B)** monoid homomorphisms are expressed as arrows.

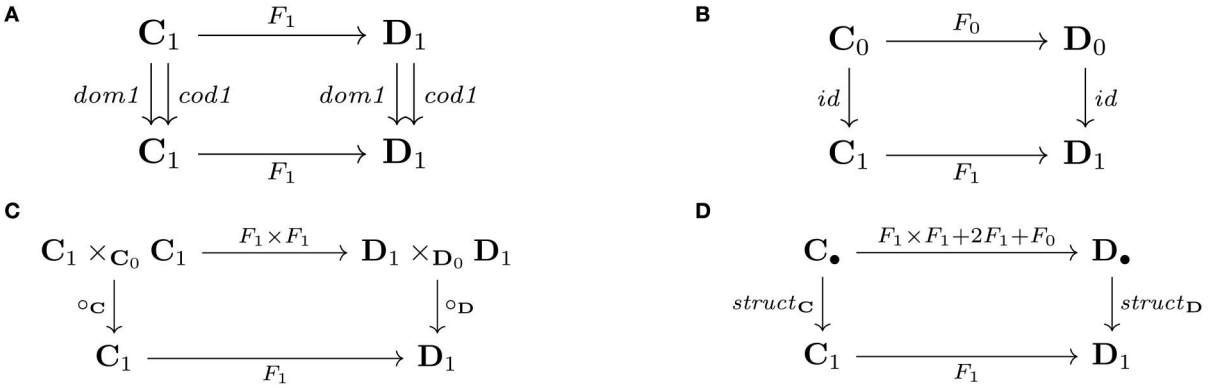

FIGURE 7
Functors as commutative squares preserving **(A)** domains/codomains, **(B)** identities, and **(C)** compositions, that is, **(D)** category structure.

Since functors are expressed as commutative squares, they also compose in a way that satisfies associativity and unity conditions, that is, categories, functors, and functor composition constitute another category (Remark 13).

Remark 13. *Commutative squares compose as commutative squares, hence composition of functors $F : \boldsymbol{C} \to \boldsymbol{D}$ and $G : \boldsymbol{D} \to \boldsymbol{E}$ is a functor, $G \circ F : \boldsymbol{C} \to \boldsymbol{E}$. So, (small) categories, functors, and functor composition constitute another category,* ***Cat****. (In this context, small means that the collections of objects and arrows are sets, not proper classes.)*

A simple way of composing representations is to take products. Products are constructed from product functors (Example 14). For instance, assuming a set of symbols for *John* and *Mary*, $S = \{\mathsf{John}, \mathsf{Mary}\}$, the product functor constructs the set of symbol pairs $S \times S = \{(\mathsf{John}, \mathsf{John}), (\mathsf{John}, \mathsf{Mary}), \ldots\}$, which can be used for to represent the *John is to the left of Mary* situation.

Example 14 (product/coproduct functors). *The product and coproduct functors send pairs of objects and arrows to their products and coproducts, respectively, that is, $\Pi : (A, B) \mapsto A \times B, (f, g) \mapsto f \times g$ and $\amalg : (A, B) \mapsto A + B, (f, g) \mapsto f + g$. In the category of sets and functions,* ***Set****, the product of two sets is (designated as) their Cartesian product and the product of two functions $f : A \to C$ and $g : B \to D$ is the product function, $f \times g : (a, b) \mapsto (f(a), g(b))$. The coproduct of two sets is (designated as) their disjoint union, and the coproduct of two functions f and g is the coproduct function, $f + g : A + B \to C$; $(1, a) \mapsto f(a), (2, b) \mapsto g(b)$. The coproduct for sets and functions acts like alternation: if a is an element from set A, then apply function $f : A \to C$, otherwise apply function $g : B \to C$.*

A product functor takes objects in a category **C** to (product) objects in **C**. In this way, constructions can be reiterated to generate compositional representations whose constituents are themselves compositional as is supposed for a language of thought. For instance, *Sue is to the left of John who is to the left of Mary* may be represented by a construction that involves the pair $(\mathsf{Sue}, (\mathsf{John}, \mathsf{Mary}))$ capturing a hierarchical relationship, that is, *Sue* is to the left of both *John* and *Mary*. A general approach to this situation involves *monoidal categories* (Mac Lane, 1998; Leinster, 2014), that is, a generalization of monoid where the set is replaced with a category and the binary operation with a functor. In this context, the functor is a kind of *tensor product* (Example 15), which affords a categorical (symbolic-vectorial) form of compositional representations (Coecke et al.,

2010). Connectionist (neural network) models employing *tensor product networks* (Smolensky, 1990) essentially involve such functors for a category of vectors spaces and linear functions.

Example 15 (tensor product). *A tensor product is a functor of the form* $F : \mathbf{C} \times \mathbf{C} \rightarrow \mathbf{C}$*, also written* $\otimes(-,-)$ *with the action on a pair of objects* (A, B) *and a pair of arrows* (f, g) *written* $A \otimes B$ *and* $f \otimes g$*, respectively. Associativity and unity conditions need only hold up to natural isomorphism (see Definition 16), that is,* $A \otimes (B \otimes C) \cong (A \otimes B) \otimes C$ *and* $A \otimes I \cong A \cong I \otimes A$*, where* I *is a special object in* $\mathbf{C}$ *whose role is analogous to the role of the unit element, e, of a monoid. For instance, product and coproduct functors (Example 14) are tensor products: in regard to* ***Set****, now seen as a monoidal category, the Cartesian product and disjoint union with any one-element set and the empty set as the unit objects, respectively.*

Entities and their representations are also naturally regarded as residing in different domains and so involve functors between different categories. For instance, the aforementioned notion of a physical instantiation mapping of symbolic expressions to brain states is likened to a functor that preserves syntactic relations as relations between brain states, that is, a mapping F that satisfies the equality $F[P\&Q] = B(F[P], F[Q])$, where B is a function combining the corresponding brain states for expressions P and Q (see Fodor and Pylyshyn, 1988, footnote 9). This condition compares with the composition condition for functors, $F_1(g \circ_{\mathbf{C}} f) = F_1(g) \circ_{\mathbf{D}} F_1(f)$, where & corresponds to composition operation $\circ_{\mathbf{C}}$ in domain category $\mathbf{C}$ and B corresponds to composition operation $\circ_{\mathbf{D}}$ in codomain category $\mathbf{D}$. The expressions P and Q and their corresponding brain states, $F[P]$ and $F[Q]$, are arrows in their respective categories. Alternatively, the expressions and brain states can be regarded as objects in a monoidal category, whence the mapping is seen as a *monoidal functor* (Mac Lane, 1998), that is, a functor preserving the structure of a monoidal category.

Much more can be said about a functorial approach to cognitive representation by specializing to functors with additional properties. For instance, an *adjoint functor* (Mac Lane, 1998; Leinster, 2014) is a functor $F : \mathbf{C} \rightarrow \mathbf{D}$ that comes with an opposing functor $G : \mathbf{D} \rightarrow \mathbf{C}$ acting as a pseudo-inverse in the sense that the composition $G \circ F$ sends objects and arrows in $\mathbf{C}$ to objects and arrows in $\mathbf{C}$ that are closely related but not necessarily the same as the original objects and arrows. (Likewise, the composition $F \circ G$ sends objects and arrows in $\mathbf{D}$ to closely related objects and arrows in $\mathbf{D}$.) That relationship is a *natural transformation*, which we turn to next (Section 2.3). This form of bidirectionality has applications, for example, in regard to the round-trip relationship between states of the world and brains states (see Ellerman, 2016; Awodey and Heller, 2020). Another example of this adjoint situation in the context of cognitive representations and processes involves *presheaves* (Mac Lane and Moerdijk, 1992) that are set-valued functors on topological spaces as categories, where the round-trip relationship acts like a generalization process in the sense that learning a training set extends (generalizes) to correct responses on a test set (Phillips, 2018, 2020). Adjoint situations involve additional category theory concepts, that is, *universal constructions* (Mac Lane, 1998; Leinster, 2014), that go beyond the detailed comparisons presented here (see Section 3).

2.3. Natural transformations and comparison

Cognitive processes are generally regarded as computational processes over (cognitive) representations. Natural transformations are maps between functors, functors were interpreted as representations, so natural transformations can be interpreted as computational processes on representations. As we shall see in this section, however, natural transformations also afford a closely related interpretation as comparisons of representations.

Definition 16 (natural transformation, isomorphism). *Suppose* $F, G : \mathbf{C} \rightarrow \mathbf{D}$ *are a pair of functors. A natural transformation* $\eta : F \dot{\rightarrow} G$ *is a family of* $\mathbf{D}$*-arrows* $\{\eta_A : F(A) \rightarrow G(A) | A \in C_0\}$ *such that* $G(f) \circ \eta_A = \eta_B \circ F(f)$ *for all arrows* $f : A \rightarrow B$ *in* $\mathbf{C}$*, which can be expressed as a commutative square (Figure 2). Arrow* η_A *is called the component of* η *at* A*. If every component* η_A *is an isomorphism—arrow* $f : A \rightarrow B$ *is an isomorphism if there exists an arrow* $g : B \rightarrow A$ *such that* $f \circ g = 1_B$ *and* $g \circ f = 1_A$*—then the transformation is called a natural isomorphism.*

The definition says that a natural transformation is composed of a family of maps from the image of one functor to the image of another functor. Hence, a natural transformation can be interpreted as a computational process for transforming representations. However, the family of maps is also required to satisfy the commutativity condition involving a square of arrows. This condition means that the transformed representations must also be comparable to the original representations. Two simple examples illustrate this situation (Example 17).

Example 17 (natural projections/injections). *The projection functors send pairs of objects and arrows to their components: e.g.,* $\acute{\Pi} : (A, B) \mapsto A, (f, g) \mapsto f$*. Product, coproduct, and projection functors are related by natural transformations: e.g.,*

- *natural projection* $\acute{\pi} : \Pi \dot{\rightarrow} \acute{\Pi}$ *(Figure 8A) and*
- *natural injection* $\acute{\imath} : \acute{\Pi} \rightarrow \amalg$ *(Figure 8B).*

Projections and injections are conceptualized by analogy to instances of matrix reasoning (Figures 8C,D), which highlights

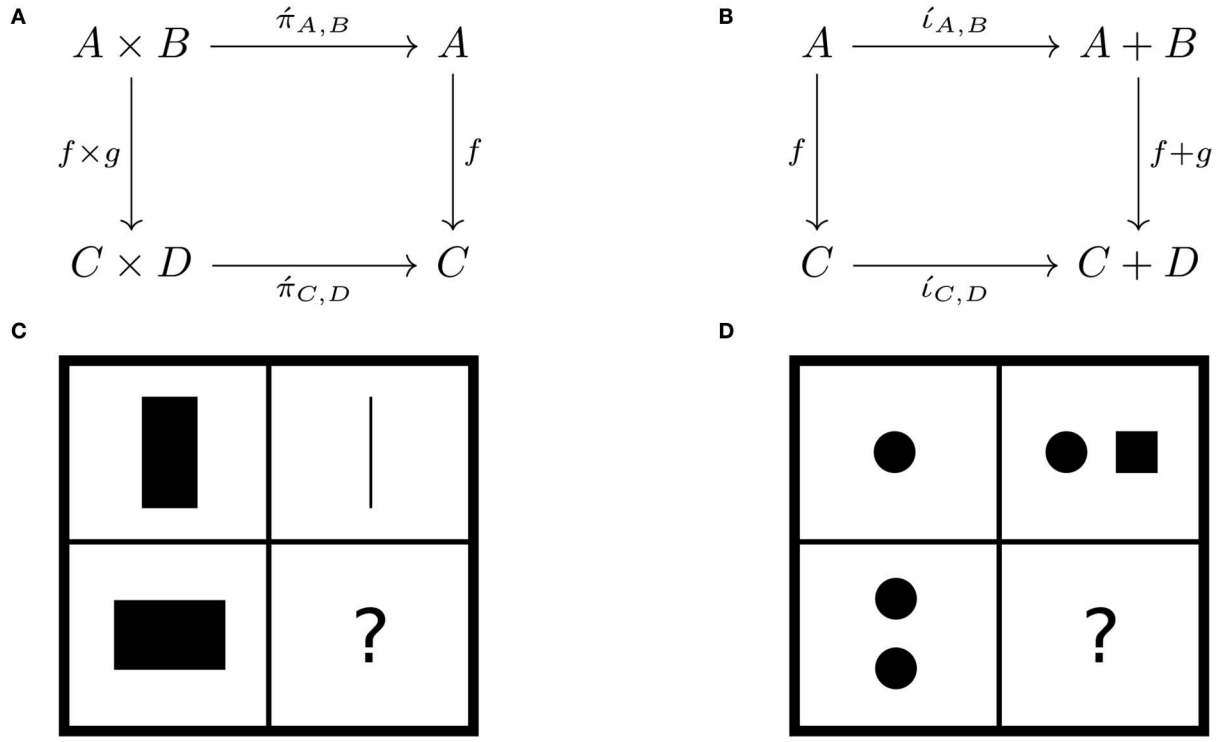

FIGURE 8
Two natural transformations: **(A)** natural projection and **(B)** natural injection as progressive matrices **(C,D)**, respectively.

the interpretation of natural transformation as a comparison of representations.

Recall (Section 1) that composition operates at all levels: composition of arrows (between objects), functors, and natural transformations. These levels also pertain to another notion of dimensionality that is illustrated with a series of progressive matrices (Figure 9). Identity arrows are interpreted as zero-dimensional in terms of variation, that is, the domain and codomain objects are the same object. Hence, there is a corresponding sense of dimensionality with progressive matrices in terms of stimulus variation along rows and columns of the matrix. For instance, when all cells contain the same stimulus (Figure 9A), the number of dimensions of variation is zero. This situation corresponds to an identity natural transformation on a *constant functor*, that is, a functor sending every object and every arrow to the same object and its identity arrow. A single dimension of stimulus variation along columns (Figure 9B) corresponds to an identity natural transformation on a functor that is not a constant functor. Alternatively, a single dimension of stimulus variation along rows corresponds to a non-identity natural transformation on a constant functor. Two dimensions of stimulus variation, along rows and columns (Figure 9C), correspond to a (non-identity) natural transformation of (non-constant) functors. Three dimensions of stimulus variation, one dimension along columns and two dimensions along rows (Figure 9D), correspond to composition of a functor with a natural transformation, which is a natural transformation. Composition of a functor with a natural transformation is also called *star composition* to distinguish this form from composition of natural transformations and composition of functors.

Natural transformations mark a significant departure from functors in terms of dimensionality. A functor also involves a square of arrows (see Figure 2A), hence looks like a natural transformation. Indeed, every functor $F : \mathbf{C} \to \mathbf{D}$ is equivalent to a natural transformation between functors into $\mathbf{C}+\mathbf{D}$ comparing the domain and codomain of F (Figure 10), that is, the natural transformation $\phi_F : \iota_0 \stackrel{\cdot}{\to} \iota_1 \circ F$, where $\iota_0 : A \mapsto (0, A), f \mapsto (0, f)$ and $\iota_1 \circ F : A \mapsto (1, F(A)), f \mapsto (1, F(f))$. However, the injection of the domain is effectively an identity map, hence involves no variation of objects and arrows along this direction. Thus, in general, natural transformations involve an extra dimension of variation compared to functors.

The examples of progressive matrices involve geometrical shapes, which highlight another aspect of category theory—generality. Up to this point, the examples have been about orders or simple algebraic structures. However, categories also exist for other kinds of structures including topological, metrical, and geometrical spaces. For instance, the category **Vec** has

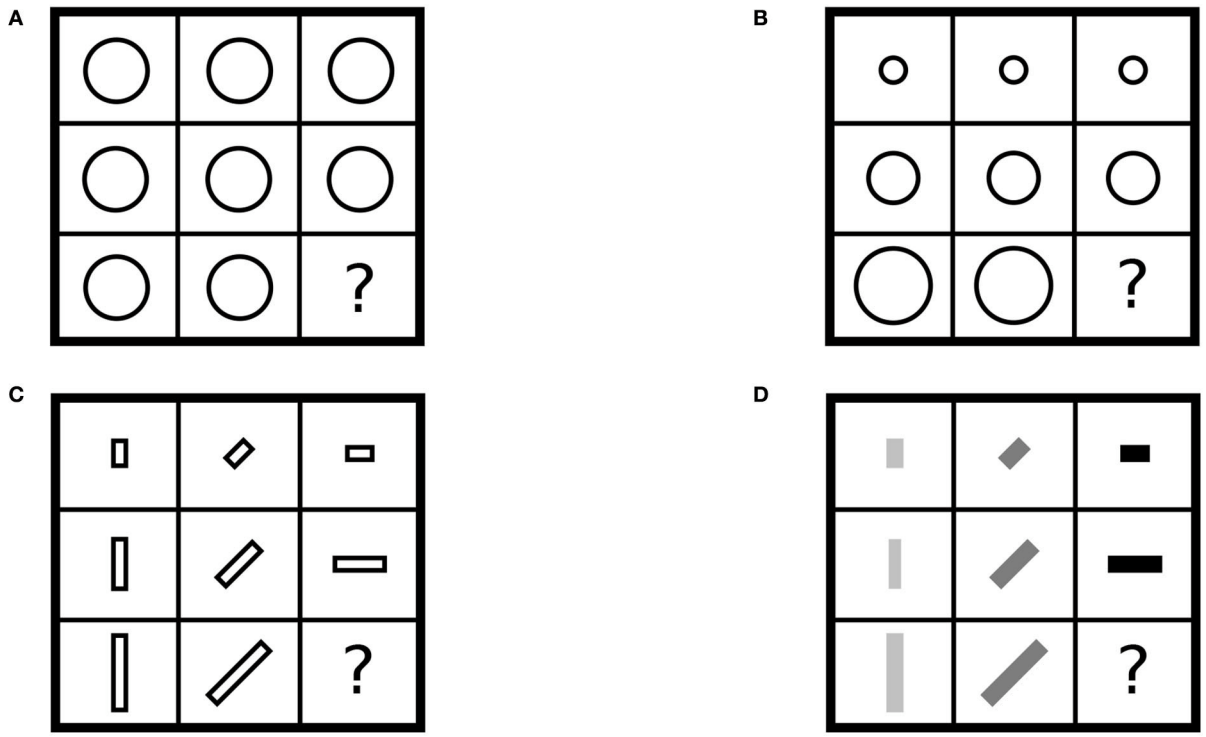

FIGURE 9
Dimensions of variation in matrix reasoning: **(A)** zero, **(B)** one, **(C)** two, and **(D)** three.

vector spaces for objects and linear maps for arrows. Thus, we can think of the shapes in terms of vector spaces and the transformations as linear maps such as dilation and rotation. The correspondence to categories, functors, and natural transformations is analogous, that is, the columns are functors from some category into **Vec** with shapes corresponding to vectors and arrows to linear transformations; likewise for the components of a natural transformation.

3. Discussion

What is category theory to cognitive science? In a broad sense, category theory like cognitive science is about the (re-)representation and comparison of compositional structure *via* maps that preserve that structure. Category theory may appear as a bewildering array of abstract definitions, examples, and theorems, but a substantial amount of this theory organizes around a simple idea, that is, the (typed) commutative square, in various contexts and forms, that comports with the view of cognition as a system of computational processes over (cognitive) representations, that is, some version of a representational/computational theory of mind (Wilson, 1999).

Representational (mental) states are supposed to capture the structure of the world, or other mental states by structural correspondence, that is, the relations between entities in the domain being represented are supposed to map in some consistent way to relations between entities in the domain of representations (cf. Frege's *compositionality principle*, or Gentner's *structure mapping theory*). A commutative square embodies this idea in geometric-algebraic form, that is, a "vertical" arrow is a structural relation in one domain that is transported, or transformed to a vertical arrow as a structural relation in another (possibly identical) domain by "horizontal" arrows that maintain structural consistency—the action on some (re-)representation is essentially the same as a (re-)representation of an action. The conditions for being a category, functor, or natural transformation mean that not any square of typed entities constitutes a commutative square. Category theory provides a vast formal generalization of this simple idea in a way that is unique among formal frameworks.

Category theory differs from other frameworks in regard to the notion of compositionality. The classical form of compositionality turns on the notion of *tokening*, that is, the representations of an entity's constituents are tokened (instantiated, activated, or inscribed) whenever the entity's representation is tokened (Fodor and Pylyshyn, 1988). For instance, a classical representation of *red circle* involves the tokening of corresponding representations for constituents *red* and *circle*. A composition operation need not token arrows

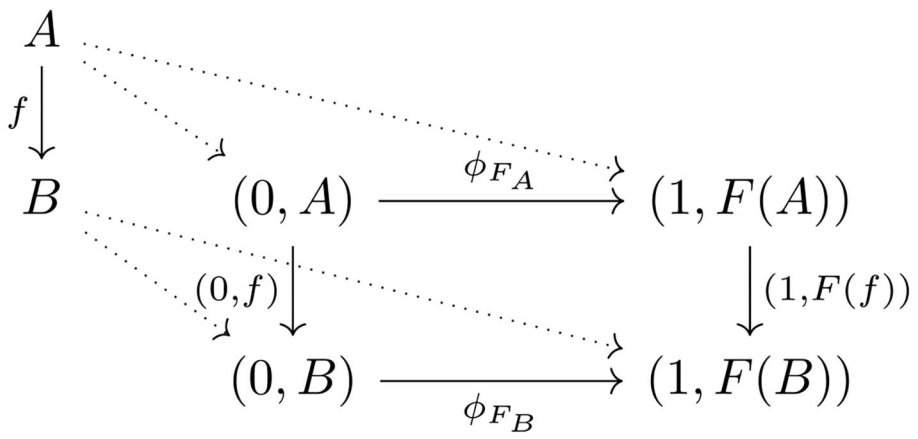

FIGURE 10
Functor as natural transformation.

in this way, as illustrated in the case of monoids where the composition of arrows 1 and 2 is their addition 3, because composition is a function that sends a pair of arrows to an arrow. A category theory analog of tokening is the *free monoid* on a set of characters (alphabet) $\mathcal{A}$ which consists of the set of all strings of zero or more characters, $\mathcal{A}^*$, that is, the monoid $(\mathcal{A}^*, \cdot, \epsilon)$, where $\cdot$ is concatenation and ϵ is the empty (length zero) string: e.g., composition of strings (arrows) "a" and "b" is the string "ba" (Walters, 1991).

Other (related) notions of compositionality were alluded to in the form of products and coproducts of objects and arrows: e.g., the product functor applied to two objects A and B constructs the product object $A \times B$. What that means precisely depends on the categories involved and the specific choice of object. For instance, in **Set**, a product of sets and A and B is designated as their Cartesian product, also written $A \times B$, and two functions affording the retrieval of the constituents of each pair (a, b), that is, a and b. There may be more than one product for a pair of objects, for example, $B \times A$ also constitutes a product for A and B. These examples seem to suggest that this form of categorical compositionality is a version of classical compositionality in that the constituent objects A and B are "tokened" with the composite object $A \times B$. However, other products of sets exist that do not involve constructing pairs of constituent elements. This difference is starker in the context of ordered sets as categories where the product of two objects A and B is the *infimum*, that is, the greatest object less than both A and B. For example, in an ordered set as a category of *divisors* (e.g., $2 \rightarrow 6$ says *two is a divisor of six*), the categorical product of 12 and 9 is the greatest common divisor, that is, 3. Category theory provides a precisely defined and vastly generalized notion of this form of composition that involves the concepts of *limit* and *universal construction* (Mac Lane, 1998; Leinster, 2014). A limit is a kind of "optimal" construction, that is, the best one can do in the given context and also a universal construction expressing a property common to all instances in that context (see Phillips, 2021a, for an introduction in the context of cognition). Though not taken up here, *systematicity* (Fodor and Pylyshyn, 1988), that is, the co-existence of cognitive abilities, is seen as a consequence of a (categorical) universal construction (Phillips and Wilson, 2010).

The import of category theory concepts to cognitive science does not end there. Base concepts of category, functor, and natural transformation constitute the starting point for category theory like the concepts of composition, representation and computation (comparison) constitute a starting point for a science of cognition. Natural transformations afford inference as do comparisons of representations. An analogy to perception provides an illustration. Inferring distance to an object is afforded by comparison of images obtained from binocular vision. A category theory analog is the *Tannakian reconstruction theorem* affording reconstruction of an object from its category of representations (NLAB, 2010). Underpinning this theorem is the *Yoneda lemma* (Mac Lane, 1998; Leinster, 2014), a fundamental result in category theory (Riehl, 2016) that relates the structure of an object to its afferent/efferent arrows. The reconstruction theorem was applied to the *relational schema induction* paradigm (Halford et al., 1998a) to account for learning transfer (Phillips, 2021b). Reconstruction involves a higher dimensional form of comparison, called a *dinatural transformation* between *bifunctors*, which are functors on two categories (cf. bivariate function). Inference is afforded by a *duality*—two opposing relations—between schemas (more generally algebras) and their representations that is analogous to a well-known duality in geometry: Two points determine a line; dually, two (intersecting) lines determine a point. For comparison, source and detection determine line of sight; dually, intersecting lines of sight determine the source. Reconstruction

involves computing the *end* of a (bi)functor (Mac Lane, 1998), that is, the "best" (universal) higher dimensional comparison.

Higher dimensional constructions relate to higher cognitive capacities, as alluded to in the comparison of matrix reasoning examples. Matrix reasoning is generally more difficult when the stimuli vary along more dimensions (Carpenter et al., 1990; Kroger et al., 2002) corresponding to a notion of cognitive complexity as the number of dimensions of task variation (Halford et al., 1998b), which has been interpreted in terms of categorical products (Phillips et al., 2009). Note that category, functor and natural transformation also have corresponding geometrical interpretations as points, lines, and sheets, hence as zero-dimensional, one-dimensional, and two-dimensional objects, respectively. Indeed, these concepts unify in higher category theory as instances of *n cells*: e.g., a *2-category* that has (small) categories as 0 cells (objects), functors as 1 cells (arrows between 0 cells), and natural transformations as 2 cells (arrows between 1 cells). This notion of dimensionality is akin to the *order* of a function or relation—a second-order relation is a relation between (first-order) relations—as another measure of cognitive complexity (Zelazo and Frye, 1998).

The import of category theory to cognitive science may seem obscured by the many technical details that could be relegated to a secondary source. However, what counts as the primary focus of attention depends on the task at hand. Moving to higher constructions is not about simply affording more general generalizations (abstractions), but rather reconciling a seemingly opposed need for concreteness. This situation is exemplified with the associativity and unity conditions for composition, which seem innocuous when viewed as counterparts in elementary algebra, but are of fundamental importance to category theory and by comparison cognitive science. For instance, the associativity condition is recovered from a natural transformation between *hom-functors* (Mac Lane, 1998; Leinster, 2014), which determine the afferent/efferent arrows for a given object in the category of interest. The clockwise and anticlockwise traversals of the commutative square for the natural transformation between hom-functors correspond to the two alternative orders of composition, that is, $h \circ (g \circ f)$ vs. $(h \circ g) \circ f$. A cognitive counterpart concerns dual-route theories (Kahneman, 2011; Evans and Stanovich, 2013), simply illustrated here by the relative difficulty of calculating $(13 \times 47) \times 0$ vs. $13 \times (47 \times 0)$, which affords a category theory way of thinking about and empirically investigating dual-route cognitive processes (Phillips et al., 2016, 2017). The identity arrows for the unity condition are also of fundamental importance despite appearances as the "do nothing" arrows. Tannakian reconstruction depends on the Yoneda lemma which in turn depends on having identity arrows. The identity arrows, appearing as elements in a hom-set of arrows for the proof of the lemma (see, e.g., Leinster, 2014), essentially ground the abstraction, cf. Frege's *principle of contextuality*, whereby the meaning of a word is determined in the context of other words (Janssen, 2001). For instance, the Yoneda lemma is seen as a way of assessing a subjective experience by its relationships to other subjective experiences (Tsuchiya and Saigo, 2021). This Yoneda/Tannaka perspective suggests an addendum to the import of category theory to cognitive science as self-referential comparison.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This work was supported by a Japanese Society for the Promotion of Science Grant-in-Aid (20H05710).

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Zelazo, P. D., and Frye, D. (1998). Cognitive complexity and control: Ii. the development of executive function in childhood. *Curr. Direct. Psychol. Sci.* 7, 121–126. doi: 10.1111/1467-8721.ep1077476

Awodey, S., and Heller, M. (2020). The homunculus brain and categorical logic. *Philos. Problems Sci.* 69, 253–280. doi: 10.48550/arXiv.1903.03424

Carpenter, P. A., Just, M. A., and Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices test. *Psychol. Rev.* 97, 404–431. doi: 10.1037/0033-295X.97.3.404

Coecke, B., Sadrzadeh, M., and Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *ArXiv preprint arXiv:1003.4394*. doi: 10.48550/arXiv.1003.4394

Eilenberg, S., and Mac Lane, S. (1945). General theory of natural equivalences. *Trans. Am. Math. Soc.* 58, 231–294. doi: 10.2307/1990284

Ellerman, D. (2016). On adjoint and brain functors. *Axiomathes* 26, 41–61. doi: 10.1007/s10516-015-9278-7

Evans, J. S. B. T., and Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Fodor, J. A. (1975). *The Language of Thought*. New York, NY: Crowell.

Fodor, J. A. (1985). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind* 94, 76–100. doi: 10.1093/mind/XCIV.373.76

Fodor, J. A., and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71. doi: 10.1016/0010-0277(88)90031-5

Fong, B., and Spivak, D. I. (2019). *An Invitation to Applied Category Theory: Seven Sketches in Compositionality*. Cambridge, UK: Cambridge University Press. doi: 10.1017/9781108668804

Fuyama, M., Saigo, H., and Takahashi, T. (2020). A category theoretic approach to metaphor comprehension: theory of indeterminate natural transformation. *Biosystems* 197, 104213. doi: 10.1016/j.biosystems.2020.104213

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cogn. Sci.* 7, 155–170. doi: 10.1016/S0364-0213(83)80009-3

Gentner, D., and Forbus, K. D. (2011). Computational models of analogy. *Wiley Interdisc. Rev. Cogn. Sci.* 2, 266–276. doi: 10.1002/wcs.105

Halford, G. S., Bain, J. D., Maybery, M. T., and Andrews, G. (1998a). Induction of relational schemas: common processes in reasoning and complex learning. *Cogn. Psychol.* 35, 201–245. doi: 10.1006/cogp.1998.0679

Halford, G. S., Wilson, W. H., and Phillips, S. (1998b). Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *Behav. Brain Sci.* 21, 803–831. doi: 10.1017/S0140525X98001769

Janssen, T. (2001). Frege, contextuality and compositionality. *J. Logic Lang. Inf.* 10, 115–136. doi: 10.1023/A:1026542332224

Janssen, T. M. V. (1997). "Compositionality," in *The Handbook of Logic and Language* (Oxford, UK: North-Holland), 417–473.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus & Giroux. doi: 10.1111/j.1539-6975.2012.01494.x

Kroger, J. K., Sabb, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., and Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cereb. Cortex* 12, 477–485. doi: 10.1093/cercor/12.5.477

Kubota, A., Hori, H., Naruse, M., and Akiba, F. (2017). A new kind of aesthetics—the mathematical structure of the aesthetic. *Philosophies* 2, 14. doi: 10.3390/philosophies2030014

Lakoff, G., and Núñez, R. (2000). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics Into Being*. New York, NY: Basic Books.

Lawvere, F. W., and Schanuel, S. H. (2009). *Conceptual Mathematics: A First Introduction to Categories*. Cambridge, UK: Cambridge University Press.

Leinster, T. (2014). *Basic Category Theory, vol. 143 of Cambridge Studies in Advanced Mathematics*. Cambridge, UK: Cambridge University Press.

Mac Lane, S. (1986). *Mathematics: Form and Function*. New York, NY: Springer-Verlag.

Mac Lane, S. (1998). "Categories for the working mathematicia," in *Graduate Texts in Mathematics, 2nd Edn* (New York, NY: Springer).

Mac Lane, S. and Moerdijk, I. (1992). "Sheaves in geometry and logic: A first introduction to topos theory," in *Graduate Texts in Mathematics* (New York, NY: Springer).

Mannone, M., and Favali, F. (2019). "Categories, musical instruments, and drawings: a unification dream," in *Mathematics and Computation in Music. MCM 2019. Lecture Notes in Computer Science, vol. 11502 of Lecture Notes in Artificial Intelligence*, eds. M. Montiel, F. Gomez-Martin, and O. A. Agustín-Aquino (Cham: Springer), 59–72.

NLAB (2010). *Tannaka duality*. Available online at: https://ncatlab.org/nlab/show/Tannaka+duality.

Phillips, S. (2018). Going beyond the data as the patching (sheaving) of local knowledge. *Front. Psychol.* 9, 1926. doi: 10.3389/fpsyg.2018.0192

Phillips, S. (2020). Sheaving—a universal construction for semantic compositionality. *Philos. Trans. R. Soc. B* 375, 20190303. doi: 10.1098/rstb.2019.0303

Phillips, S. (2021a). A category theory principle for cognitive science: Cognition as universal construction. *Cogn. Stud. Bulle. Jpn. Cogn. Sci. Soc.* 28, 11–24

Phillips, S. (2021b). A reconstruction theory of relational schema induction. *PLoS Comput. Biol.* 17, e1008641. doi: 10.1371/journal.pcbi.1008641

Phillips, S., Takeda, Y., and Sugimoto, F. (2016). Why are there failures of systematicity? the empirical costs and benefits of inducing universal constructions. *Front. Psychol.* 7, 1310. doi: 10.3389/fpsyg.2016.01310

Phillips, S., Takeda, Y., and Sugimoto, F. (2017). Dual-routes and the cost of determining least-costs. *Front. Psychol.* 8, 1943. doi: 10.3389/fpsyg.2017.01943

Phillips, S., and Wilson, W. H. (2010). Categorial compositionality: a category theory explanation for the systematicity of human cognition. *PLoS Comput. Biol.* 6, e1000858. doi: 10.1371/journal.pcbi.1000858

Phillips, S., Wilson, W. H., and Halford, G. S. (2009). What do transitive inference and class inclusion have in common? categorical (co)products and cognitive development. *PLoS Comput. Biol.* 5, e1000599. doi: 10.1371/journal.pcbi.1000599

Raven, J., Raven, J. C., and Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales: General Overview*. San Antonio, TX: NCS Pearson.

Riehl, E. (2016). *Category Theory in Context*. New York, NY: Dover.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intell.* 46, 159–216. doi: 10.1016/0004-3702(90)90007-M

Tsuchiya, N., and Saigo, H. (2021). A relational approach to consciousness: categories of level and contents of consciousness. *Neurosci. Consciousn.* 7, 1–13. doi: 10.1093/nc/niab034

Walters, R. F. C. (1991). *Categories and computer science, vol. 28 of Cambridge Computer Science Texts*. (Cambridge, UK: Cambridge University Press).

Wilson, R. A. (1999). "Philosophy," in *The MIT Encyclopedia of the Cognitive Sciences*, eds. R. A. Wilson and F. C. Keil (Cambridge, MA: MIT Press). xv–xxxvii.