A unified, contemporary approach to teaching energy in introductory physics

Ruth Chabay\textsuperscript{a)} and Bruce Sherwood\textsuperscript{b)}

\textit{Department of Physics, University of North Texas, Denton, Texas 76203}

Aaron Titus\textsuperscript{c)}

\textit{Department of Physics, High Point University, High Point, North Carolina 27268}

(Received 25 November 2018; accepted 8 May 2019)

Students in introductory physics courses frequently struggle with various aspects of energy analyses. We argue that inconsistencies and errors in the traditional treatment of energy contribute to these student confusions. As an alternative to creating multiple activities to ameliorate student difficulties, we describe a more coherent, contemporary approach to the teaching of energy that offers students a principled way to avoid confusions. © 2019 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.1119/1.5109519

I. INTRODUCTION

Inconsistencies are common in the theoretical framework for energy in the traditional introductory calculus-based physics curriculum, and these inconsistencies may cause difficulties for students who are learning about energy.\textsuperscript{1,2} The work-energy theorem, which is derivable from Newton’s second law, is often not distinguished from the true energy equation; as a result, students may choose an inappropriate equation to start a problem.\textsuperscript{3} The terms “conserved” and “constant” are often used interchangeably; as a result, students may believe that conservation of energy is not general.\textsuperscript{4} The choice of the system, though recognized to be important in free-body analyses, is often not given adequate attention in energy analyses, where potential energy may be erroneously ascribed to a single object rather than to a pair of interacting objects within a system; as a result, students may double-count work and \( \Delta U \) terms. Binding energies, which are positive numbers, are not carefully defined as the energy required to dissociate a system, and so students are often puzzled that \( K + U \) for a bound system (such as a planet orbiting a star) is a negative number. An arbitrary constant is often added to potential energy (although in the larger relativistic context this is not possible); this can aggravate confusion about the signs of potential energy terms. Paradoxes occur when friction (a dissipative phenomenon involving deformable systems) is invoked while modeling objects as point particles (which are not deformable and have no internal degrees of freedom). Despite the fact that friction forces are, at a fundamental level, conservative interatomic electric forces, they are often labeled “non-conservative,” leaving students wondering if electric forces must also be “non-conservative.” Rather than attempting to add activities that give students practice in dealing with such confusions, we advocate a restructuring of the energy component of the introductory physics curriculum in a manner that is coherent, consistent, and contemporary, and which empowers students to analyze interesting phenomena such as fission and fusion simply by applying fundamental principles.

II. TOO MANY ENERGY EQUATIONS

For a student, it can be bewildering that there seem to be several different fundamental energy equations: the “work-energy theorem,” the “first law of thermodynamics” used in the thermal physics context, and the essentially nameless true energy equation used in the mechanics course. By being explicit about whether a system is modeled as a point particle or an extended system with internal degrees of freedom, it is possible to reduce this list to just one equation.

A. The work-energy theorem

Energy is often introduced in the context of the “work-energy theorem” which is actually a path integral of Newton’s second law and fundamentally deals with momentum, not energy. Starting from Newton’s second law applied to a multiparticle system, \( \frac{dp_{cm}}{dt} = \mathbf{F}_{\text{net}} \), and integrating through the displacement of the center of mass, one obtains the nonrelativistic form of the work-energy theorem.

\[
\text{The work–energy theorem: } \Delta \left( \frac{1}{2} M_{\text{total}} v_{cm}^2 \right) = \int \mathbf{F}_{\text{net}} \cdot d\mathbf{r}_{cm}. \tag{1}
\]

This equation tells us that the change in translational kinetic energy of a system is equal to the net force applied to the system acting through the displacement of the center of mass of the system. One can derive the work-energy theorem by starting from the components of \( \frac{dp_{cm}}{dt} = \mathbf{F}_{\text{net}} \) and obtain the following:

\[
\Delta \left( \frac{1}{2} M_{\text{total}} v_{cm,x}^2 \right) = \int F_{\text{net},x} dx_{cm}, \tag{2}
\]

\[
\Delta \left( \frac{1}{2} M_{\text{total}} v_{cm,y}^2 \right) = \int F_{\text{net},y} dy_{cm}, \tag{3}
\]

\[
\Delta \left( \frac{1}{2} M_{\text{total}} v_{cm,z}^2 \right) = \int F_{\text{net},z} dz_{cm}. \tag{4}
\]

The sum of these three equations yields the work-energy theorem. The fact that there are three separately valid component equations makes it clear that the work-energy theorem, despite its name, is basically a momentum equation, not an energy equation. Further evidence for this point of view is that there is no internal energy term in the work-energy theorem. It only predicts the change in the translational kinetic energy of the system, independent of rotational, vibrational,
or thermal energy changes. Note that in these equations, the relevant displacement is that of the center of mass, which may or may not be equal to the displacement of the point of action of each individual force on the system.

B. The energy principle

There does not seem to be a widely accepted name for the true energy equation in the context of the introductory mechanics course; this equation relates the change in the total energy of a system to the energy inputs to the system from the surroundings. We choose to give the true energy equation the name used by Fred Reif, the “Energy Principle.” It cannot be derived from \( \frac{d\vec{p}}{dt} = \vec{F}_{\text{net}} \) but rather is the fruit of nearly 200 years of experimental and theoretical work. It has the following form, where \( E_{\text{sys}} \) is the energy of the chosen system, and energy transfers can be in the form of mechanical work, thermal energy transfer (microscopic work), radiation, etc.:

\[
\Delta E_{\text{sys}} = \int (\vec{F}_i \cdot d\vec{r}_i) + \text{other energy inputs from surroundings.} \tag{5}
\]

In the Energy Principle, the mechanical work done on the system by the surroundings is calculated by summing the work done by each individual force, \( \int \vec{F}_i \cdot d\vec{r}_i \). If the displacements of the individual forces are not the same as \( d\vec{r}_{cm} \), the right side of the work-energy theorem, which is sometimes called “pseudowork,” will not be equal to the work term in the Energy Principle. In 2008, John Jewett published a useful five-part tutorial on these matters in The Physics Teacher.

Figure 1 illustrates a simple example: pull on the left and right ends of a spring with your left and right hands, with each hand applying a force of the same magnitude. Applying the Energy Principle to the extended system of the spring, your left hand does positive work and your right hand also does positive work; the net input of energy has the effect of increasing the potential energy of the spring (which is a deformable multiparticle system and can have potential energy associated with the interactions of the atoms within the system).

In contrast, applying the work-energy theorem (thus modeling the system as a point particle, Fig. 2), we find that the net force on the system is zero and acts through a zero displacement of the center of mass, which correctly yields zero change in the translational kinetic energy.

Similarly, consider pushing to the right at the top of a disk rotating on a horizontal low-friction axle and simultaneously pushing with the same force to the left on the bottom of the disk. Both hands do positive work to increase the rotational kinetic energy of the disk modeled as an extended system, but the net force and the displacement of the center of mass are both zero, in agreement with the fact that the translational kinetic energy of the rotationally accelerating disk remains zero.

C. Conservation of energy

The most general statement of the Energy Principle emphasizes that energy is always a conserved quantity, where \( E_{\text{surr}} \) is the energy of the surroundings

\[
\Delta E_{\text{sys}} + \Delta E_{\text{surr}} = 0. \tag{6}
\]

This is a proper statement of conservation of energy, and one says that energy is a conserved quantity. It is not uncommon, but incorrect, to speak of the energy in an inelastic collision as not being “conserved” when in fact what is meant is that the kinetic energy did not remain constant; energy is always a conserved quantity.

In a situation where \( \Delta E_{\text{sys}} = 0 \), this is simply the energy principle applied to an isolated system. If one applies the energy principle only to isolated systems, students may confuse the more general conservation of energy with the specific case where the energy of a system is constant. When students study special relativity in later physics courses, it will be useful to distinguish between the separate ideas conserved, constant, and invariant.

D. Point-particle vs extended-system models

Beginning students have difficulty making fine distinctions, and it is unfortunate that the work-energy theorem and the Energy Principle may look rather similar to the student. A more vivid way of keeping them distinct is to apply the same Energy Principle to two different models of a system: first, the actual multiparticle “extended” system, and second, a point-particle system located initially at the center of mass of the extended system, with the same mass and same center-of-mass velocity and acted on by the same net force as the extended system. The point-particle system’s motion will be exactly the same as the motion of the center of mass of the extended system, but the only change in its energy will be translational kinetic energy, and because all of the forces act through the displacement of the point particle, the work done on the point particle may be different from the mechanical work done on the extended system, as in the case of the spring discussed above.

The kinetic energy of a multiparticle system can usefully be written as a sum of translational kinetic energy and kinetic energy relative to the center of mass, including rotational and vibrational kinetic energy

\[
K_{\text{total}} = K_{\text{translation}} + K_{\text{relative}}. \tag{7}
\]

Changes in translational kinetic energy can be calculated by applying the Energy Principle to the system modeled as a point-particle, thereby evaluating one of the terms in the Energy Principle that is applied to the extended system

\[
\Delta K_{\text{translation}} = \int \vec{F}_{\text{net}} \cdot d\vec{r}_{cm}. \tag{8}
\]
By learning to apply the energy principle separately to the point-particle model and the extended-system model, students gain experience in a valuable aspect of doing physics-modeling. They learn to define the model, identify assumptions inherent in the model, and recognize limitations of the model.\textsuperscript{12}

III. CHOICE OF THE SYSTEM AND POTENTIAL ENERGY

All approaches to problems involving freebody diagrams emphasize the importance of careful definition of the chosen system, but often there is less care when doing energy analyses, where the choice of the system is just as important. The key issue is that potential energy is a kind of internal energy—it is a property of an interaction between two objects within a system; a system of a single particle does not have potential energy.

Consider the case of a rock falling a distance $h$ near the surface of the Earth. If we choose the rock alone as the system, the Earth is part of the surroundings and can do work on the rock (Fig. 3). In this case, the work done by the Earth is $+mgh$, and so

$$\Delta K_{\text{sys}} = W = +mgh.$$  \hfill (9)

Alternatively, if we choose the rock plus the Earth as the system, negligible work is done by the surroundings, and energy flows between kinetic energy and potential energy within the system (Fig. 4). The potential energy change in the (rock + Earth) system is

$$\Delta U = \Delta (mgy) = -mgh,$$ \hfill (10)

and so

$$\Delta K_{\text{sys}} + -mgh = 0.$$ \hfill (11)

Either choice of the system will give the appropriate result for $\Delta K_{\text{sys}}$ (the kinetic energy change of the Earth is negligible). However, it is not uncommon to say that the rock has potential energy $mgy$ and that its kinetic energy increases as its potential energy decreases, $\Delta K + \Delta (mgy) = 0$. Given this confusion, it is quite reasonable for the careful student who knows about potential energy and about work to include both entities and say $\Delta K + \Delta (mgy) = W = mgh$, thereby double-counting and getting an incorrect result.

A clearer approach is to emphasize that a single object (the rock in this case) does not have potential energy. Potential energy is a function of the relative positions of two interacting objects and might well be called “configurational energy.”

The concept of the system can be further elaborated by discussing how the form of the Energy Principle changes for different choices of the system (and surroundings), using as an example a woman lifting and accelerating a barbell, choosing the system to be the woman, barbell, and Earth, then the barbell alone, and then the barbell and the Earth. Similarly, it is useful to show how the form of the Energy Principle changes when the system is viewed in a moving reference system, in which case each individual energy and work term changes yet the Energy Principle is still valid.

It can be very useful to engage students in making graphs of kinetic, potential, and total mechanical energy in systems that include gravitational orbits and in hanging spring-mass systems in arbitrary 3D motion. In the case of orbits and electric interactions, it is useful to have students plot energies not only as a function of separation distance, which is the most common presentation, but also as a function of time (Fig. 5). Graphs of $K$, $U$, and $K + U$ as a function of time make clear the flow of energy within a multi-object system. Graphs of energy vs. separation distance make it easy for students to identify bound states as states with negative potential energy and total mechanical energy.

A natural extension is to introduce the basic aspects of discrete energy levels in microscopic systems, with emphasis on graphs indicating energy levels and transitions between levels. Working with energy graphs and discrete energy levels in the introductory course provides a foundation for the understanding of such graphs in the later modern physics and quantum mechanics courses.

IV. ENERGY IN THE RELATIVISTIC CONTEXT

There are several pedagogical advantages to embedding introductory mechanics in the larger relativistic framework, by explicitly including the concept of rest energy. It makes it possible to engage students in analyzing interesting particle reactions such as neutron decay, fission reactions, and fusion reactions (providing an interesting context for the application of electric potential energy). It can also help students become comfortable with the negative sign of potential energy terms.

Einstein’s theory predicts and the results of experiments confirm that the total energy of a single particle is $E_{\text{particle}} = \gamma mc^2$, where $\gamma = 1 / \sqrt{1 - v^2/c^2}$. This particle energy can be written as $\gamma mc^2 = mc^2 + K$, separating it into rest energy and kinetic energy.
Consider a system of two objects with the same mass \( M \) (for example, two identical stars or an electron and a positron) that are at rest so far apart that their interaction is negligible. Since their kinetic energy is zero, the total energy of this system must be \( 2Mc^2 \). Since the system includes two objects, we can write the total energy of the system as the sum of the rest energies, kinetic energies, and the potential energy associated with the interaction

\[
E_{\text{total}} = 2Mc^2 + K_1 + K_2 + U.
\]

In order for the total energy of the system to be equal to \( 2Mc^2 \), the potential energy \( U \) must be zero at very large separations; one cannot add an arbitrary constant. Furthermore, the energy and momentum four-vector will not transform properly unless in this situation \( U = 0 \). That which was relative before 1905 is now absolute, which may be conceptually simpler for students.\(^{13}\)

The total energy of a multi-particle system of mass \( M \) whose center of mass is at rest is \( Mc^2 \). This rest energy can also be written as the sum of the masses and kinetic energies of all the system’s constituents plus the sum of all the pair-wise potential energy terms. The mass \( M \) can be less than the sum of the individual masses of the unbound particles, in which case we define the binding energy as the amount of energy needed to dissociate the system into its constituents. For example, the mass of the bound \( O_2 \) molecule is very slightly less than the mass of two oxygen atoms, while in nuclear physics the mass difference between bound and unbound nucleons can be sizable. Although the mass difference is tiny in chemistry, it is conceptually useful to understand binding in these absolute terms, which helps avoid a common student misconception of thinking of binding energy as something “stored in chemical bonds” rather than being associated with a negative mass difference. Moreover, the total energy of a bound system, including the rest energy, is always positive, which can help students accept situations in which potential energy terms are negative.

V. DISSIPATIVE INTERACTIONS

For systems that can be adequately modeled as point particles (that is, there is no change in internal energy), one can demonstrate path independence, but in the case of dissipative forces such as sliding friction or air resistance, path independence does not hold. These forces are sometimes called “nonconservative” forces, which is an unfortunate term, because these interatomic forces are electric interactions, and electric forces obey path independence. The issue is not the kind of force but rather the kind of system. A sliding block is a multiparticle system whose internal energy increases as a result of inelastic collisions of the block with the surface.\(^6\) The phenomenon of sliding friction is fundamentally entropic, associated with the huge number of atomic degrees of freedom of the sliding block.

Of course analysis of the point-particle version of the block (which is equivalent to applying the work-energy theorem) can be used to determine the change in the translational kinetic energy, but this analysis ignores the obvious fact that the block’s internal energy increases, as indicated by a temperature increase.

VI. INTERNAL ENERGY

Students are sometimes confused about the difference between internal energy and other forms of energy. In fact, internal energy is a convenient catch-all term that denotes “forms of energy that we choose not to analyze in detail at the moment,” and not a fundamental form of energy.\(^{14}\) The clearest and most parsimonious view is that there are only three distinct categories of system energy: rest energy, kinetic energy, and potential energy (or equivalently, in more advanced courses, field energy). However, it is often convenient to define “internal energy” as, for example, the sum of the kinetic and electric potential energy associated with the thermal motions of all the atoms in a solid object. Internal energy need not refer solely to atomic-level kinetic and potential energy; the energy associated with the rotation or vibration of components of a macroscopic system can also
be termed “internal energy” if we are interested only in the translational kinetic energy associated with motion of the center of mass of the system.

A microscopic viewpoint is useful in clarifying these ideas and can help students see that thermal energy is not a different form of energy but simply kinetic and potential energy at an atomic level. This makes it possible to see thermal physics and mechanics as closely related rather than as completely distinct.\textsuperscript{15,16}

\section*{VII. WHY FIELDS?}

A puzzle associated with potential energy provides a motivation for mentioning the field concept, which may be of interest to some students. Consider two stars with equal mass $M$ at rest far from each other. To analyze the energetics of their approach, a good choice of system is the two stars, with the assumption that the surroundings are empty. Then, $\Delta E_1 + \Delta E_2 + \Delta U + \Delta E_{\text{surr}} = 0$. There is nothing in the surroundings, so $\Delta E_{\text{surr}} = 0$, and the kinetic energy of both stars increases as $U = -\frac{GM^2}{r}$ decreases.

However, our conclusion should not depend on our choice of system, so we next analyze a system consisting solely of star 1, where star 2 is the only entity in the surroundings: $\Delta E_1 = W$, and the kinetic energy of star 1 increases due to the positive work done by star 2 (there is no potential energy term for this single-object choice of system). It must also be true that $\Delta E_1 + \Delta E_{\text{surr}} = 0$. The kinetic energy of star 1 is increasing, so the energy of the surroundings must be decreasing, yet the surroundings seem to consist solely of star 2, whose kinetic energy is increasing not decreasing.

Evidently something is missing from this energy analysis. To be able to analyze this situation fully, we need to introduce the abstract idea of a “field,” which is the subject of the second-semester E&M course, where one sees that seemingly empty space can contain energy and momentum. Despite our lack of understanding, the power of the Energy Principle is such that we can calculate how much energy loss there is in the “gravitational field.” The change in the energy of the surroundings is $\Delta E_{\text{surr}} = -\Delta E_1$, with part of $\Delta E_{\text{surr}}$ being $\Delta E_2$, which is equal to $\Delta E_1$ for this symmetric situation. Hence, the change in the field energy is $-\Delta E_1 - \Delta E_2$. The analysis of the two-star system led to the same result for $U$: $\Delta U = -\Delta E_1 - \Delta E_2$, so we get the correct result using potential energy with the two-star system. However, the difficulty in accounting for the energy in the surroundings suggests that one must in some analyses include field energy.\textsuperscript{17}

\section*{VIII. ASSESSMENT}

Some aspects of our approach to energy may seem unusually advanced for the calculus-based introductory physics course taken by science and engineering students. However, this approach, incorporated into the introductory calculus-based physics textbook \textit{Matter & Interactions},\textsuperscript{18} is used with many different student populations at a variety of different kinds of institutions: large and medium-sized universities, four-year liberal arts colleges, community colleges, and a few high schools. The fact that some of these topics are typically encountered only in later physics courses is a matter of tradition, not of necessity.

Evaluating the success of the approach requires assessment instruments. Although widely known, the Energy and Momentum Conceptual Survey (EMCS),\textsuperscript{19} because it is aligned with the traditional presentation, encourages some of the confusions discussed above. In particular, some questions confuse the terms “constant” and “conserved,” and attribute potential energy to a system of a single object. A more extensive, curriculum-specific instrument\textsuperscript{20} was developed to assess learning of the approach to energy discussed in this paper and has been used to assess student learning in courses that use this approach. Although the validity and reliability of this instrument are clearly established, it is not typically used in traditional introductory physics courses because about half of the items either assess understanding of the framework discussed in this paper, or refer to topics not covered in the traditional course (such as rest energy).

\section*{IX. CONCLUSIONS}

We have highlighted what we consider to be significant problems with the treatment of many energy topics in the traditional calculus-based introductory physics course, and we have presented ways in which these problems can be addressed. Issues include confusion of the work-energy theorem with the true energy equation, the need to be more careful in specifying the system of interest, the fact that single particles do not have potential energy (which is a property of pairs of interacting particles), confusion between “constant” and “conserved” quantities, the value of embedding energy in the larger relativistic context, and the fact that friction forces are not “nonconservative” (being interatomic electric forces). We argue that the more coherent approach to energy discussed in this paper is not only accessible to students at the introductory level, but can help eliminate some of the confusions that students encounter in the traditional approach to energy at this level.

\section*{ACKNOWLEDGMENTS}

The authors thank Gregg Franklin and Curtis Meyer for helping them to think deeply about energy. The authors also thank Fred Reif for emphasizing the role of the three fundamental principles of mechanics, which he called the Momentum Principle (Newton’s second law), the Energy Principle, and the Angular Momentum Principle.

\begin{thebibliography}{99}
\item \textsuperscript{a}Electronic mail: Ruth.Chabay@unt.edu
\item \textsuperscript{b}Electronic mail: Bruce.Sherwood@unt.edu
\item \textsuperscript{c}Electronic mail: attitus@highpoint.edu
\item Beth A. Lindsey, Paula R. L. Heron, and Peter S. Shaffer, “Student ability to apply the concepts of work and energy to extended systems,” Am. J. Phys. 77, 999–1009 (2009).
\item Beth A. Lindsey, Paula R. L. Heron, and Peter S. Shaffer, “Student understanding of energy: Difficulties related to systems,” Am. J. Phys. 80, 154–163 (2012).
\end{thebibliography}
Froment Motor

I photographed this unusual form of electric motor at St. Patrick’s College in Maynooth, Ireland in the fall of 1999. It was designed by the French engineer, Paul-Gustav Froment (1815-1865) who in 1844 devised an electric motor that was one of the first used for industrial purposes. In his design, electromagnets are energized to pull in iron bars mounted on a revolving cage. Once the iron bar is level with the electromagnet, the current is cut off until the next iron bar is in range. A commutator is used to make and then break the current to the electromagnet. (Picture and text by Thomas B. Greenslade, Jr., Kenyon College)