International Journal of Technology and Design Education (2021) 31:981–1000
https://doi.org/10.1007/s10798-020-09579-w

Practice before theory? An approach for testing sequencing effects in pedagogical design

Sukumar Natarajan[1] · Nick McCullen[1] · Steve Lo[1] · David Coley[1] · Omaimah Ali Arja[2,3] · Francis Moran[1]

Accepted: 19 March 2020 / Published online: 19 April 2020

Abstract
Engineering is a practical discipline, dedicated to the solution of problems through the sound application of principles derived from the natural sciences and mathematics. Engineering pedagogy has therefore to balance the need for learners to gain a deep understanding of the theoretical basis of the problem domain whilst grasping its practical implications. However, little is known as to the most effective *sequence of delivery*: is it better to begin with theory and build up to practice or vice versa? Here, we present the idea of testing this through a carefully designed pedagogical experiment. We begin by discussing the issues around the creation of a pedagogical experiment to answer such a question, and define the nature and scope of such experiments. We then create a formal framework within which such questions can be tested and present an experiment in the domain of architectural engineering that pilots this new approach. Finally, we discuss the utility of using such a framework to lead evidence-based discussions of pedagogical practice within the engineering education literature, and conclude that similar experiments could be, and should be, completed by other teams wanting to examine delivery order or other binary choice situations.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10798-020-09579-w) contains supplementary material, which is available to authorized users.

✉ Sukumar Natarajan
s.natarajan@bath.ac.uk

Nick McCullen
N.J.McCullen@bath.ac.uk

Steve Lo
s.n.g.lo@bath.ac.uk

David Coley
d.a.coley@bath.ac.uk

Omaimah Ali Arja
omaimah.ali@gju.edu.jo

[1] Department of Architecture and Civil Engineering, University of Bath, Bath BA2 7AY, UK

[2] School of Architecture and Built Environment, German Jordanian University, Madaba Street, Amman 11180, Jordan

[3] Faculty of Architecture and Design, Middle East University, Airport Street, Amman 11180, Jordan

Keywords Theory versus practice · Sequencing · Problem based learning

Introduction

As a practical discipline, engineering requires an understanding of sound scientific principles and the ability to apply these in 'messy' settings to solve real-world problems. From a pedagogical standpoint, this means that a course of learning must include both Theory and Practice to be effective in enabling students to learn. Recognition of this has led to the development of new pedagogical constructs such as Action-Based-Learning (ABL) and Problem-Based Learning (PBL) as alternatives to the more traditional Information-Based-learning (IBL).

When teaching, it is obvious that the overall balance of theory and practice—or indeed other elements such as quantity and type of assessment—in a programme of learning can affect student engagement, motivation and eventually the attainment of the stated learning outcomes. For example, Törley emphasises the importance of "seeing" and "hearing" in the learning process to aid in a process of de-abstracting (our term) principles, such that they can be *applied* (Törley 2014). In engineering education, in particular, there is a growing recognition of practice as a concept that is generally under-valued. However, there is no single agreed definition in the literature of what constitutes "practice".

Drawing primarily on the influential theories of Schatzki (Schatzki 2012) and others such as Gherardi (2009) and Fenwick (Fenwick and Edwards 2010; Fenwick 2012; Fenwick et al. 2015), Hager et al. (2012) suggest a series of principles within which we can understand practice:

1. *Knowing-in-practice:* embodies the idea that learning goes beyond, or in some sense cannot be conflated with, the acquisition or transfer of knowledge. It is a complex process that involves the body, language and interaction with many "others" including people, processes and organisations. In our work, we understand this as developing an 'intuition of theory'.
2. *Socio-materiality:* suggests that practice is constituted within an arrangement of material and non-human actors in space and time. For example, the particular arrangement, type and location of equipment in an industrial laboratory will influence practice.
3. *Embodied and relational:* practices are *embodied*—i.e. not just as ideas within the head but through the use of the body itself, through the use of the hand in creation or during walk-through inspections etc. *Relationality* of practice comes from the observation that the process of creation results from, or is highly influenced by, relationships with both human and non-human entities. For example, products or services can be seen as being co-produced with clients or vendors.
4. *Contextual evolution:* suggests that practice exists and evolves in well-defined contexts that have social and historical influences. Within engineering, in particular, history and context can play a significant role such as the use of the British Thermal Unit (BTUs) as the unit of energy in North America, in contrast to the use of the kilowatt-hour (kWh) in Europe. This is purely a matter of practice and contextual evolution as the quantity being measured is simply energy (as heat).
5. *Emergence:* borrows from complexity theory where the ultimate state of a system cannot be specified or known in advance, but rather emerges through the complex interaction of elements within the system. Practice is said to be emergent in much the same way

> since the way in which it changes is subject to indeterminable externalities or other perturbations.

Indeed, Reich et al. use the above principles in support of their argument that understanding practice in engineering requires a fundamentally different approach; one that recognises that the engineering workplace is more global, complex and interdisciplinary (Reich et al. 2015).

This type of understanding of practice is relatively modern. Concerns in the 1970s and 1980s around an over-emphasis on theory in traditional engineering curricula, eventually led to a different type of resurgence in practice based pedagogy. We now understand this as the use of "capstone projects" or Problem Based Learning to introduce practice based elements in learning (Liebman 1989). Here, practice is defined more narrowly as the process of applying knowledge towards the solution of a real-world problem. The idea being that through the application of knowledge towards a problem within the sandboxed constraints of a learning environment, students will be able to "learn to design". In fact, the disciplines of civil engineering and architecture almost exclusively rely on this type of project driven learning as a means to train graduates due to the size of the structures or systems involved (Dutson et al. 1997). In other areas such as chemical engineering or mechanical engineering, physical apparatus or prototypes become feasible, and are often used. Indeed, it is probably true that the definition of a "practice element" in engineering education that an educator today is most likely to recognise is one of problem-based or project-based learning. Despite their similarities, these approaches must not be conflated. A "capstone" project derives its name from the decorative capstone used to finish a building. Hence, it is, by definition, an integrative element that will occur towards the *end* of a programme. For example, at the University of Bath, architecture and engineering students collaborate on a major design project that consumes the better part of one semester of their final year. Problem Based Learning on the other hand can occur at any time within the curriculum. Engineering educators have placed more emphasis on science as a theory (Banios 1992). The shift towards a more theoretical approach in the engineering curriculum has resulted in far less experienced graduates in the practice of engineering and design than those in previous years (Liebman 1989).

Our work is agnostic towards the definition of both theory and practice in engineering education. We implicitly recognise that both theory and practice are evolving constructs that are often highly individual to a person or an organisation. Rather, our concern is to understand whether there is a *sequencing effect*, i.e. whether it is better to start with theory and build up to practice or vice versa. There is little evidence in the literature that this issue has been debated or systematically tested. Indeed, in a large systematic review of over 100 papers by Dutson et al. (1997) on the use of capstone projects, the question of whether to use a *capstone* or its antithesis a *foundation stone*, is hardly addressed. More broadly, the term "sequence", when it does occur in the literature, usually pertains to a complete thread of activities (where practice, if it occurs, is implicitly capstoned), rather than a structured debate about the positive or negative effect of sequencing as an act of pedagogical practice (Gander et al. 1994). Here, we argue that an evidence-led approach might usefully contribute to answering this question. However, we make no claim to either certitude or completeness, and recognise that pedagogical practice itself is broad and many-valued.

Testing for sequencing effects requires careful study design, and the definition of what constitutes a meaningful difference. Our overall aim here, is to present a framework for conducting a pedagogical experiment to answer this question.

Designing a pedagogical experiment

Creating meaningful and reliable experimental conditions in education can be difficult due to the need to ensure fairness to learners, consistency of experimental conditions and comparability of results. When testing sequencing effects, each of these issues can present unique problems:

- Fairness: Given that most pedagogical experiments are undertaken under 'live' settings, the experimental design needs to be created so as to generate little or no difference between groups in terms of overall student experience.
- Consistency: This is a narrower requirement than that of fairness since it implies the minimisation of differences between the groups other than the variable of interest, in this case *sequencing*. Differences can arise from a range of factors including differences in (1) venue, (2) tutors and (3) pedagogical materials. Ensuring consistency of some elements (e.g. venue, tutors) can be easier when the groups are not concurrent, for example, between years, but this can make direct comparisons difficult.
- Comparability: Groups can only be compared when they have been created through a process of *random* allocation. Hence, it can be difficult to undertake such experiments in situations where random allocation is not possible, such as group design projects where students often choose their partners. However, as we shall show in "A pilot experiment in architectural engineering" section, this is achievable to a certain extent with planning and foresight.

In addition to the above issues, a key question is the definition of success or failure of such an experiment. For example, one could hypothesise that the sequencing effect is likely to be subtle, so how big a difference between groups should the threshold for success or failure be set to? What does this imply for sample sizes? Crucially, what effect should one measure? How is it to be measured? We discuss an approach to resolving these issues through the creation of a simple framework in the next section.

A framework for undertaking experiments

We identify the following key steps needed to undertake a pedagogical experiment:

Step O1 Define:

(a) Measurement metric used to judge success/failure.
(b) How the success/failure of the experiment will be judged, including an appropriate analysis method.
(c) Format of the experiment.
(d) Learning Outcomes (LOs), separated by the sequencing effect being tested. In our case, one set of LOs for theory and one for practice.

Step O2 Ethics:

(a) Consider the ethical implications of undertaking the experiment. Particularly, the extent to which the purpose of the research purpose is elaborated to the subjects (i.e. students), and whether the effects of observation (i.e. Hawthorne effect) or participation (e.g. Foucault's Panopticon) may play a role in the results (see, for example, Hill 2017).
(b) Ensure participant data is anonymised but can be traced (e.g. for score normalisation under 'before' and 'after' conditions).
(c) Obtain ethical approval from an appropriate body (e.g. departmental ethics procedure).
(d) Obtain prior informed consent from the participants.

Step O3 Undertake the experiment, ensuring:

(a) An appropriate randomization procedure has been put in place.
(b) Fairness, consistency and comparability are not compromised.
(c) Participants are traceable (i.e. scores from the same participant can be tracked) but not identifiable (e.g. by using a suitable anonymization procedure).

Step O4 Obtain:

(a) Data pertaining to the measurement metric.
(b) Anonymised student feedback using an appropriately tailored question set, including free text responses where appropriate. Feedback must be sought on all aspects of the experiment, including the LOs, tutors, format and overall experience.

Step O5 Analyse using the stated method, after

(a) Suitable data management for any missing records, or incomplete data.
(b) Check for any 'ceiling effect' in an after test, if used (see "Defining: a measurement metric" section).

Diagrammatically, this results in the framework shown in Fig. 1. While much of **Step O2** to **Step O5** can be inductively inferred through our demonstration in "A pilot experiment in architectural engineering" section, or will be unique to the conditions of each organisation or experiment, there are aspects of **Step O1** that merit further exploration. We discuss each of these below.

Defining: a measurement metric

An holistic assessment of success or failure of a pedagogical experiment will need to use three key metrics: (1) in-session engagement, (2) student satisfaction and (3) formal attainment (assessed formatively or summatively). Elements of engagement include collaboration, project orientation, and authentic focus, usually measured through a process of observation (Kearsley and Shneiderman 1998). Hence, while an important line of evidence, it does not lend itself easily to quantitative judgement of whether learning objectives have been met.

Student satisfaction can be measured using a range of metrics (e.g., see "A pilot experiment in architectural engineering" section), and are often offered in standardised form in

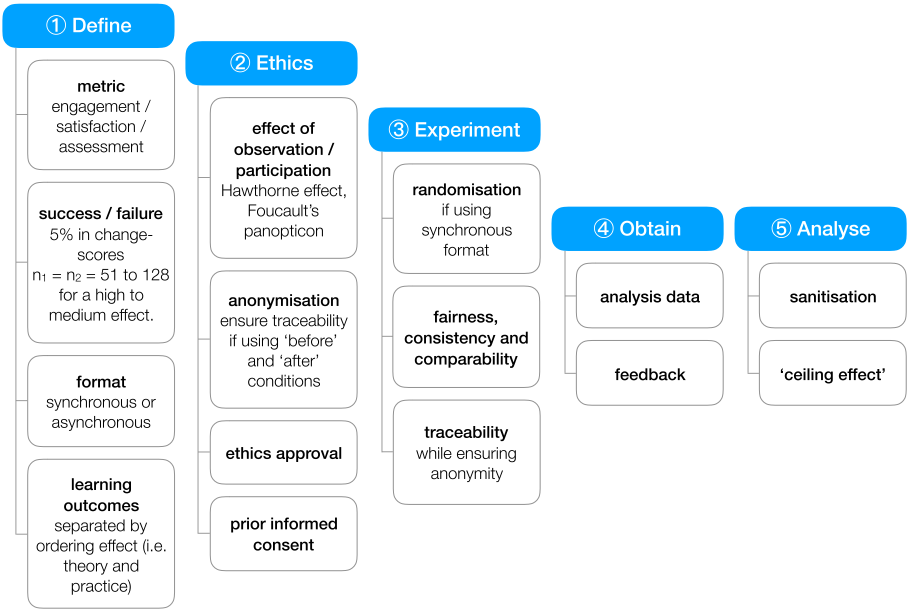

① Define
metric engagement / satisfaction / assessment
success / failure 5% in change-scores n₁ = n₂ = 51 to 128 for a high to medium effect.
format synchronous or asynchronous
learning outcomes separated by ordering effect (i.e. theory and practice)
② Ethics
effect of observation / participation Hawthorne effect, Foucault's panopticon
anonymisation ensure traceability if using 'before' and 'after' conditions
ethics approval
prior informed consent
③ Experiment
randomisation if using synchronous format
fairness, consistency and comparability
traceability while ensuring anonymity
④ Obtain
analysis data
feedback
⑤ Analyse
sanitisation
'ceiling effect'

Fig. 1 A framework for undertaking a pedagogical experiment for testing the sequencing effect of practical and theoretical elements in a programme of learning. When applied, this should be informed by the evaluator's own knowledge and practice of teaching

many institutions. This is usually quantified using ordinal data and can be useful in comparing outcomes between groups, including the attainment of learning objectives. However, these data are sensitive to the obtained return rate. Anecdotal evidence from our own institutions suggest average rates of 30–40% and no more than 50%. Hence deciding success or failure using purely satisfaction metrics exposes the analysis to the risk of low return rates.

Formative or summative assessment hence provides the most reliable means of testing success or failure. 'Before' and 'after' measures can be obtained to measure increase in attainment by either repeating the assessment or creating formally equivalent assessments. When using assessments, it is important to ensure that (1) the assessment sufficiently covers all aspects of the programme, particularly those being tested (e.g., theoretical and practical elements) (2) the assessment is properly graded in difficulty to capture a wide range of attainments and (3) when using the same assessment in 'before' and 'after' conditions, that there is sufficient room for the grades to grow in response to the learning activity and thus avoid any 'ceiling effect'[1] that could negatively affect the analysis of results.

[1] The 'ceiling effect' refers to the idea that where the same test is applied across two levels (e.g. 'pre' and 'post' groups) there must be sufficient space within the scoring system for the scores to "grow". For example, if the average scores in a 'pre' test are high (e.g. 80%) then the ability for scores to grow sufficiently in response to the 'post' test may be constrained, e.g. result in an average score of 100%. In such an instance, we would assume the difference in pre- and post- tests is 20% whereas, in a properly designed test, the same students might have revealed mean scores of 43% and 67% (or a difference of 24%).

Defining: success or failure

We begin with the assumption that *any* form of instruction will increase student attainment, compared to attainment in the subject prior to engagement with the programme (i.e. before being taught). In a *well taught* programme, student attainment will rise and satisfaction is likely to be high. Hence, simply comparing attainment and satisfaction between groups is unlikely to be of use when testing for a sequencing effect, as it does not account for the state of knowledge of the students prior to undertaking formal learning. Hence, the question becomes one of testing whether the *relative change* between groups is significant. That is, the change from a baseline in one group must be higher than the change from the baseline of the other group, to be treated as significant.

The definition of significance also requires some consideration. Standard inferential tests use the *p* value to determine significance, usually at the 0.05 level. That is, a test is conducted using randomly drawn samples to determine the likelihood that any observed difference is purely due to chance. If the probability is less than 5%, the result is deemed significant. However, this ignores the fact that even the most trivial difference becomes statistically significant with sufficiently large sample sizes. Hence, it is important to consider the *size* of the effect to determine true significance. For simple two group tests, Cohen's *d* (Eq. 1) is usually used with small, medium and large effects indicated by *d* values of 0.2, 0.5 and 0.8, respectively. Here, *d* values indicate the number of standard deviations by which the two means differ. A *d* of 0.5 hence implies that the two means differ by half a standard deviation. Cohen's *d* is calculated using the following expression:

$$d = \frac{\left|\bar{x}_{group\,1} - \bar{x}_{group\,2}\right|}{\sqrt{\frac{(s_{group\,1})^2+(s_{group\,2})^2}{2}}} \tag{1}$$

where d = Cohen's d; $\bar{x}$ = mean score of a group, s = standard deviation of a group's score.

Since attainment metrics are usually in the form of percentage scores, it is useful to ask what difference in percentage points between groups should be treated as significant. For example, if we discovered that teaching theory before practice produced a one percentage point higher score than teaching practice before theory, would this be sufficient for someone to alter their curriculum design? Few assessors would be able to confidently state that they are able to mark to a precision of 1%.

A change of 10% is usually considered large. For example, at our institutions, this will usually mean a jump in grade boundaries (e.g. difference between a 'merit' classification at 65% to a 'distinction' at 75%). Hence, we suggest that an appropriate standard for judging success or failure of pedagogical experiments would be between 1 and 10%, at about 5%. For example, our institutions consider a difference of 4% to be significant when comparing marks from two independent assessors for a single summative submission of high credit weighting.

The effect size calculated from a 5% difference in the means of two groups was determined over a range of different standard deviations—using the same standard deviation for both groups. As expected, the effect size implied by a particular shift in mean scores is smaller for larger standard deviations, as shown in Table 1. The values for Cohen's *d* can be seen to vary from very large values of 2.0 down to 0.3 in the range given. Even so, this lower end is between the low and medium effect sizes suggested by Cohen, and corresponds to very large standard deviations of 15%, which are unusual at our institutions. This confirms that, under normal circumstances, a 5% shift in the means of two test

Table 1 Indicative values for Cohen's *d* when the difference between sample means is 5% under a range of standard deviations, assuming equal standard deviations for both samples

	Standard deviation					
	2.5%	5.0%	7.5%	10.0%	12.5%	15.0%
Cohen's *d*	2.0	1.0	0.7	0.5	0.4	0.3

groups could be used to determine the effect of a measure. This is *provided* appropriate sample sizes and reasonably robust marking procedures that do not produce unusually high variances.

We can use this to undertake basic power analysis to determine the per group sample size for a range of effect sizes, with the pwr.t.test function in the statistical software R. Assuming conventional values for the Type I (i.e. $\alpha = 0.05$)[2] and the Type II (i.e. $\beta = 0.2$)[3] errors, we obtain $n_1 = n_2 = 51$, $n_1 = n_2 = 128$ and $n_1 = n_2 = 351$ for effect sizes of 0.8, 0.5 and 0.3, respectively.[4] In other words:

- Per group sample sizes between 51 and 128, assuming $\alpha = 0.05$ and $\beta = 0.20$, are sufficient to reach conclusions about effects down to $d = 0.5$.
- A five percent change in marks, expressed as 0.5 standard deviations (i.e. $d = 0.5$), would imply test scores with a standard deviation of 10%.

Defining: experimental format

Experiments with sequencing effects can be run either synchronously or asynchronously, each of which has strengths and weaknesses.

Synchronous experiments imply a doubling of resources such as tutor time, venue etc., because activities have to run in parallel to one another. In carefully controlled experiments, these resources will be identical to avoid the introduction of bias, which simply means that each element of the programme has to be taught twice by the same person under the same conditions. Hence, synchronous experiments are better suited to shorter formats such as "block" delivery where an entire module or unit is delivered within a concentrated length of time, usually measured in days. A common form for such block delivery at our institutions is 5 days, starting on a Monday morning and finishing by Friday evening. The key strengths of this format are that it utilises the same cohort of students (usually by randomly splitting them into equal groups) and, due to their short duration, can also be conducted outside formal curriculum or 'term time'.

Asynchronous experiments occur over a period of time and imply the utilisation of different cohorts of students to represent the experimental groups, possibly separated by a

[2] α is the probability of rejecting the null hypothesis when it is true. This is usually kept to a low value such as 5% to avoid a *false positive:* thinking we have found an effect where none exists.

[3] β is the probability of failing to reject the null hypothesis when it is false. We are usually more tolerant of this (say 20%) as we are trying to avoid a *false negative:* thinking there is no effect when, in fact, there is.

[4] For those unfamiliar with power analyses, there are several useful online resources such as the one hosted by the Comprehensive R Archive Network at https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html.

year if undertaken within formal learning. This carries the risk of the existence of differing ‘environmental’ conditions through, for example, differences in venue or tutors, changes in the wider institution or programme, etc. However, it has the benefit of deeper and longer immersion in a programme of study and much larger sample sizes as a single cohort no longer needs to be split into two groups.

Defining: learning outcomes (LOs)

Clearly defined learning outcomes are an essential tool in programme design, but also true for the design of pedagogical experiments. Only through clearly defining the LOs for each part of the experiment, can we properly create assessments to test whether the stated outcomes were attained. Without these, there is the risk of spurious inferences being drawn from the experiment. In our case, this translates to one set of LOs for theory and one for practice.

A pilot experiment in architectural engineering

Our goal in conducting this experiment is twofold: (1) to provide an *end-to-end demonstration* of the overall approach outlined earlier as a template for future studies (i.e. to examine the effect of sequencing) and (2) provide sample data that could seed future studies.

Our hypothesis is centred around the question of sequencing, i.e., whether it is better to sequence a programme of study with ‘theory before practice’ (TbP) or ‘practice before theory’ (PbT). We assess student attainment through a carefully designed test administered both before *and after* each delivery sequence (TbP/PbT). This allows us to control for any pre-existing differences between groups by only concentrating on comparing the *change* in score between the groups. We can therefore state the null (H_0) and alternate (H_A) hypotheses as:

H_0 TbP and PbT produce the same increase in ‘before’ and ‘after’ test scores.

H_A Either sequence (PbT or TbP) produces a higher increase in ‘before’ and ‘after’ test scores, than the other.

Study design

As the authors specialise in architectural engineering, particularly building thermodynamics, this was the broad area selected. The chosen topic for delivery was the design of refugee shelters, which confers the following benefits:

- A refugee shelter, as defined by the United Nations High Commissioner for Refugees (UNHCR 2016), is usually a simple single-zone (i.e. a single undivided room) building. This significantly reduces the complexity of heat and mass transfers, allowing the students to focus on key concepts.
- The above simplicity allows greater exploration of the practical aspects of specifying, building and monitoring the performance of a building. This is usually not possible in

larger buildings as both the construction and thermodynamics interact to create greater complexity.
- This is a topic of global concern given that there are an unprecedented number of displaced people at the time of writing (Fosas et al. 2018), with the major concern for aid agencies being the provision of good quality shelters (Fosas et al. 2017). Indeed, the choice of Amman as a location provided access to Jordanian students who are acutely aware of the crisis, given that Jordan is a key host country for Syrian refugees. This reduces the risk that 'topic engagement' would negatively affect the experiment.
- The authors are all involved in a major research project to develop a new science of shelter design. This has provided access to shelters of many kinds all over the world, creating an invaluable resource of practical and theoretical work on which to base the teaching.

Due to the choice of Amman as a location, constraints of time and resources implied the use of the synchronous format.

Defining 'theory'

The subject of refugee shelter design is vast as it traverses a range of engineering problems such as structural safety, thermodynamics and fire risk, socio-anthropological issues such as 'dignity' and 'agency', management issues such as logistics, cost and scaleability as well as domestic and international politics. Even within the relatively narrow confines of building thermodynamics, several interacting factors (e.g. external weather, construction, sizing and orientation, ventilation, insulation, mass), their associated uncertainties and possible scales of representation, create difficulty in the selection of a narrow but relatively "complete" set of principles to engage with, within an educational setting. Hence we took the following into consideration in determining an appropriate content strategy for the theory track:

- The key thermodynamic challenge associated with shelter design is to ensure good performance under extreme weather (e.g. the desert conditions prevalent in Jordan). Hence, indoor temperature performance under a single peak summer and peak winter day (evaluated over a typical year such as a Test Reference Year, i.e. an average year) were selected as representative of this challenge.
- Since our objective was to test the effect of content sequencing between theory and practice, it was important to ensure that the breadth and depth of the theoretical aspects were circumscribed by the practical aspects that could reasonably covered in the time available. Hence the examples used in the theory track were informed by data from a series of real shelters in Jordan.

This led to the following learning outcomes for the 'theory' track:

T-LO1 To understand the fundamentals of heat transfer through the building fabric and through ventilation.
T-LO2 To understand how thermal mass works.
T-LO3 Gain knowledge of key terms needed to quantify building thermodynamics including U-value, air change rate, time lag, decrement factor, and thermal time constant.

T-LO4 To be able to use the above terms by expressing their inter-relationships as equations and explore their effect through a simple spreadsheet.

Three bespoke learning sessions were created, one covering T-LO1 and associated T-LO3, the second covering T-LO2 and associated T-LO3, and the third covering T-LO4.

Defining 'practice'

As discussed in "Introduction" section There are several aspects to 'practice' in engineering, including:

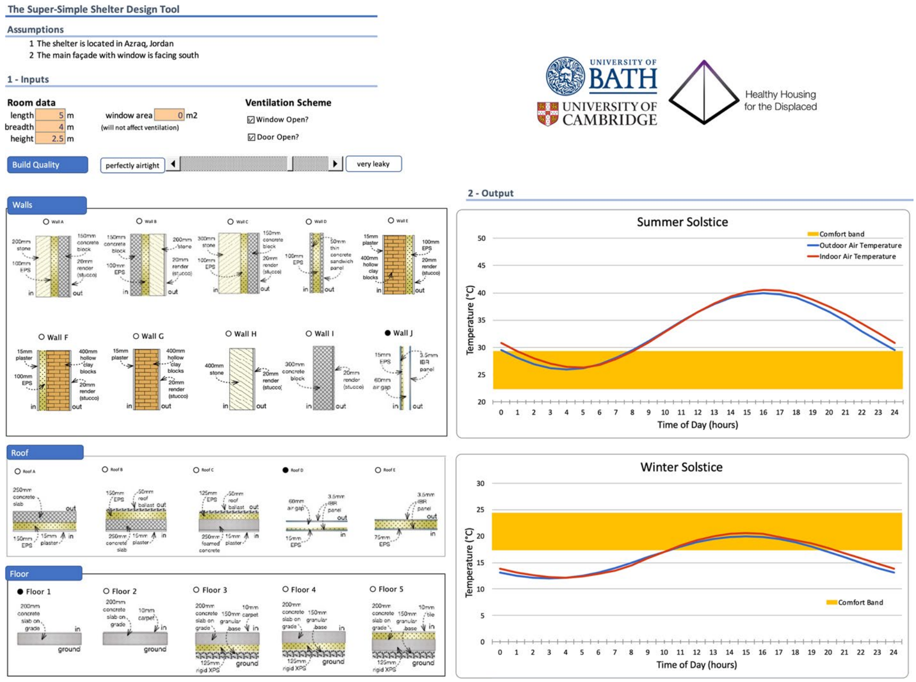
1. Developing an *intuition of theory*, that is, an understanding of the relative impact of the various aspects under the control of the designer
2. Developing an *understanding of materials, connections and assembly*
3. *Application of ideas* in a design context
4. Project management and planning
5. Prototyping
6. Lab-testing
7. Understanding contextual history
8. Developing relationships with clients or other stakeholders etc.

Of these, our 'practice' track addressed aspects (1), (2) and (3), leading to the following learning outcomes:

P-LO1 To develop an intuition of the effect of different shelter wall and roof assemblies on the likely performance on a peak summer and winter day.
P-LO2 To develop an understanding of the actual performance of different shelter wall and roof assemblies under real weather.
P-LO3 To gain an appreciation of the challenges of constructing different shelters at scale and speed.
P-LO4 To apply concepts within a design context.

To achieve P-LO1 a novel *interactive* spreadsheet tool, the Super Simple Shelter Tool (SSST) was developed that allowed students to independently explore the peak summer and peak winter impact of 250 wall and roof assemblies using simple, intuitive controls (Fig. 2). A key feature of the SSST was the presence of detailed annotation of wall and roof assemblies but absence of information relating to their thermodynamic performance (as detailed in T-LO3). This was done to ensure that the thermodynamic performance could be *inferred* rather than enumerated from the tool's prediction of summer and winter performance, to be consistent with P-LO1.

Three sessions were hence created in the 'practice' track: one covering P-LO1 via the SSST, a second covering a series of practical demonstrations derived from our own fieldwork to cover P-LO2 and P-LO3 and a third providing time for a conceptual shelter design exercise to address P-LO4.



The Super-Simple Shelter Design Tool

Assumptions

1 The shelter is located in Azraq, Jordan
2 The main façade with window is facing south

1 - Inputs

Room data
length 5 m
breadth 4 m
height 2.5 m

window area 0 m2
(will not affect ventilation)

Ventilation Scheme
☑ Window Open?
☑ Door Open?

Build Quality — perfectly airtight … very leaky

Walls: ○ Wall A ○ Wall B ○ Wall C ○ Wall D ○ Wall E ○ Wall F ○ Wall G ○ Wall H ○ Wall I ● Wall J

Roof: ○ Roof A ○ Roof B ○ Roof C ● Roof D ○ Roof E

Floor: ● Floor 1 ○ Floor 2 ○ Floor 3 ○ Floor 4 ○ Floor 5

2 - Output

Fig. 2 The Super Simple Shelter tool, developed to address P-LO1. The left hand side contains inputs including a series of wall-roof-floor combinations whose effect can be observed in peak summer and winter on the right hand side. The red and blue curves in the graphs are the external and internal temperature time series, and the orange bands provide an indication of the "ideal" comfort bands from the literature. (Color figure online)

Table 2 The format of the experiment

	Session 1	Session 2	Session 3	Session 4	Session 5
Day 1	Intro + pre-test	*T-LO4* **P-LO1**	*T-LO1 + T-LO3* **P-LO2 + P-LO3**	*T-LO2 + T-LO3* **P-LO4**	P-LO4
Day 2	**T-LO4** *P-LO1*	**T-LO1 + T-LO3** *P-LO2 + P-LO3*	**T-LO2 + T-LO3** *P-LO4*	P-LO4	Post-test + presentation

Sessions starting with a 'T' are on the theory track, and those with a 'P' on the practice track. Participants were randomly allocated to two equal groups, coded as italic (TbP group) and bold (PbT group) on Day 1, Session 1. The design brief was also introduced at the same session. All sessions were of equal length

Format

The experiment was conducted in Amman, Jordan, in January 2019, and had the following features:

- Instruction was undertaken in an intensive two-day "block" format (see Table 2), outside the normal academic calendar.

- Both PbT and TbP tracks were run in parallel, in two separate but adjoining rooms of identical size and orientation.
- Three lecturers were involved in delivery, two assigned to delivering different aspects of ‘practice’ and the other to ‘theory’. This effectively controlled for differences in delivery, as each track was always taught by the same tutor/s.
- The programme included a mixture of on-screen instruction (including pictures and video from our field work) and interactive off-screen learning, all of which was produced collaboratively.

It is noteworthy that each of the last three points are made to ensure consistency in delivery and experience across the groups.

Participants

Participation was entirely voluntary, with participants being drawn from second and third year students on a 5 year architectural degree. This ‘self-selection’ aids the experiment as it minimises the likelihood of the results being influenced through topic disengagement. Participants had not been formally taught a significant portion of the course material, though some basics had been covered within the main curriculum. Informed consent was sought from each participant and the experiment was subject to an ethical approvals process at the University of Bath. Participants were told they would be tested before and after the programme and that they would be split into two groups for simultaneous delivery. However, the explicit goals of the pedagogical experiment were not explained to minimise the Hawthorne effect and Foucault’s panopticon.

There were a total of 22 students with each being randomly allocated to a TbP Group and a PbT Group. Students were asked to put their names on the test sheet for traceability but these were subsequently anonymised during transcription. One student in the PbT group arrived late for the baseline test (hereafter ‘Pre Test’) and a different student did not complete the second test (hereafter ‘Post Test’). Hence, we obtained unequal group sizes for groups TbP (n_{TbP}= 11) and PbT (n_{PbT} = 9). Based on the power analysis in “Defining: success or failure” section, the obtained sample sizes are too small to directly answer the hypothesis. However, the obtained data can be treated as a pilot experiment that can be used to undertake more robust power analysis by using, for example, the observed group standard deviations and effect size.

Test

A test was designed to measure student learning of key concepts across both tracks. The same test was administered at the start of the experiment to produce a baseline measure of performance against which change in score could be measured for each student. The test was in Multiple Choice Question (MCQ) format, with questions of varying difficulty covering:

1. Understanding of terms and units
2. Ability to compute key quantities, given data
3. Understanding of the likely performance of different construction assemblies
4. Knowledge of the relative performance of different materials
5. An understanding of the impact of construction detailing on performance

Table 3 Test scores for pre- and post-tests for both groups (*TbP* and *PbT*)

Group	Participant ID	Pre-test	Post-test	ΔTest scores
TbP	P1	8	21	13
	P2	10	18	8
	P3	4	23	19
	P4	2	14	12
	P5	6	19	13
	P6	17	15	-2
	P7	6	16	10
	P8	10	18	8
	P9	6	12	6
	P10	6	6	0
	P11	2	12	10
	$\bar{x}_{TbP}$	18%	40%	22%
	s_{TbP}	11%	12%	15%
PbT	P12	16	21	5
	P13	25	29	4
	P14	12	29	17
	P15	4	21	17
	P16	8	23	15
	P17	10	25	15
	P18	8	16	8
	P19	0	10	10
	P20	8	17	9
	$\bar{x}_{PbT}$	25%	53%	28%
	s_{PbT}	18%	16%	13%
	p	0.27	0.05	0.18

Max score is 40. Group means are given as $\bar{x}$ and sample standard deviations as s

6. An understanding of design choices in terms of orientation and shading

Results

The complete anonymised results can be seen in Table 3. All students taking both tests completed the tests. Questions with no answers were given zero. Given that the max score under post-test is 29, we conclude there is no ceiling effect preventing scores from growing further.

The *TbP* group showed a lower mean increase in scores (20%) compared to the *PbT* (30%), so $\Delta_{means} = 10\%$. Unsurprisingly, given the sample sizes, a t test suggests $p = 0.18$ which is not significant at the 0.05 level. It is noteworthy that the test we have suggested in this paper for such experiments, i.e. the use of a *change* from baseline scores as the preferred measure ("Defining: success or failure" section), is stringent when compared to a naïve use of the t-test on the final scores themselves. For example, in our experiment, a t-test on the raw scores for the post-test (i.e. "endline" scores) where $\overline{TbP}_{post} = 40\%$ and $\overline{PbT}_{post} = 53\%$(i.e. $\Delta_{means} = 13\%$) produces $p = 0.05$, which one might be tempted to claim as significant. However, this ignores the effect of baseline knowledge which, even

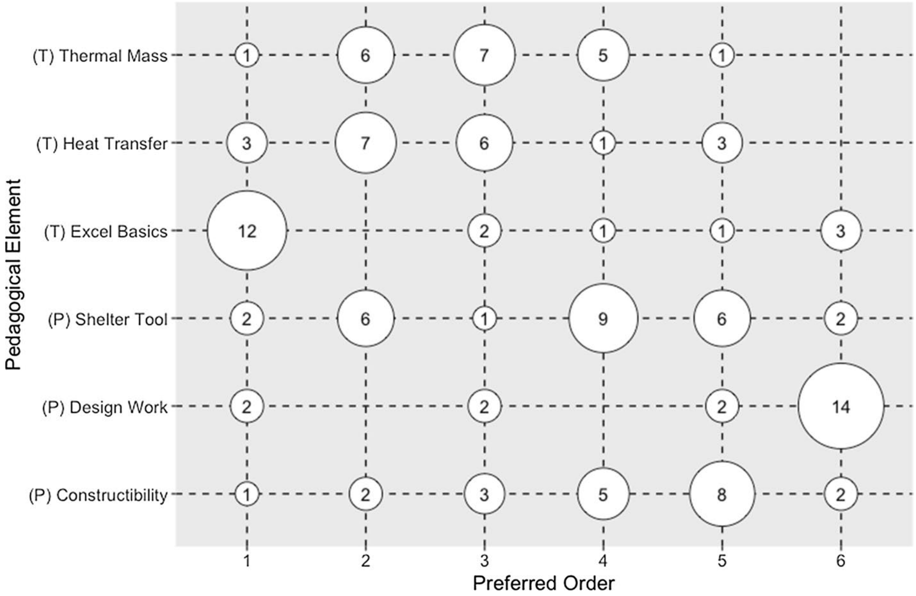


Fig. 3 Student feedback on the preferred order (x-axis) of pedagogical elements (y-axis). The top three rows are on the theory track (T) and the bottom three on the practice track (P). The size of each bubble indicates the number of students voting for a given element at a given position. The preferred order is Excel Basics → Shelter Tool → Heat Transfer → Thermal Mass → Constructability → Design Work

with random sampling, can be biased in one group compared to the other when dealing with small sample sizes. In our experiment, the higher attainment of the *PbT* group compared to the *TbP* group in the post-test was offset by their higher attainment in the pre-test ($\overline{TbP}_{pre}$ = 18% and $\overline{Pbt}_{pre}$ = 25%), reducing overall significance. Cohen's *d* for the experiment was 0.23 suggesting any effect, if it exists, may be small.

Student feedback was, in general, highly positive, with a 96% return rate (21 out of 22 students). For the fifteen questions on a 5-point scale ranging from 1 (low) to 5 (high) scores, ten questions had median scores of 5 with the remaining at 4. Since the final outcome of the programme was the design of a new shelter, we sought feedback on which part of the programme best aided this outcome. This seems to suggest a strong preference for the practice track (43%) over theory (19%) or both in equal measure (33%). This is perhaps unsurprising as design is inherently practical in nature. Students' preferred sequencing suggests a strong preference for a "Theory before Practice" sequencing (Fig. 3). These data illustrate the importance of obtaining feedback in addition to attainment data, as the choice becomes less clear if these are in conflict.

Exploring overall conceptual understanding

Ignoring the effect of sequencing allows us to gain a picture of overall performance, which was highly positive. The overall change between mean pre- and post-test scores was 24 percentage points ($score_{pre}$ = 21% and $score_{post}$ = 45%). This suggests that, and as indicated in "Defining: success or failure" section, the overall effect of the teaching is to increase student attainment. This is reflected in the fact that for 14 out of the total 17 questions in the

Table 4 Overall mean test scores by question for pre- and post-tests

Question	Pre-test (%)	Post-test (%)	Δ Test scores (%)
Q01	38	29	– 10
Q02	48	62	14
Q03	38	52	14
Q04	33	52	19
Q05	52	76	24
Q06	5	5	0
Q07	10	38	29
Q08	0	24	24
Q09	10	38	29
Q10	14	57	43
Q11	5	67	62
Q12	5	10	5
Q13	5	29	24
Q14	57	57	0
Q15	14	24	10
Q16	67	71	5
Q17	0	71%	71
Total	21	45	24

test, the mean change is + 27 percentage points. Out of the remaining three, two showed no change and one showed a drop of 10 percentage points. Here, we briefly explore pre- and post-test score changes in three categories to understand which concepts showed the *most*, *no* and *least* change. All the questions can be seen in the "Electronic Supplementary material, Appendix" and the change scores for each question can be seen in Table 4.

There were two questions that resulted in large changes in pre- and post-test scores; Q11 with a change score of 62% (pre-test baseline 5%) and Q17 with a change score of 71% (pre-test baseline 0%). It is instructive that both questions tackle essentially the same subject, the positioning of insulation with respect to thermal mass, but using different visual and written language. This links directly to the second learning outcome on the theory track T-LO2, but also to the first learning outcome on the practice track P-LO1, possibly explaining the lack of any difference between groups. These large changes for both questions suggest that the overall concept of thermal mass needing exposure to the internal air moved from being the least understood initially, to being definitively "embedded" within the students' thinking after the learning activities.

Two questions, Q6 and Q14, showed a change score of 0% albeit with very different features. Q6 addresses P-LO1 and P-LO2 tests whether students understand the effect of an air-gap as part of the insulation layer in a wall build up, the correct answer being (E). The counterintuitive result that air gaps result in lower performance than insulation of equivalent thickness due to convection, despite the lower thermal conductivity of stationary air, escaped most students. Answers were distributed across the options in the pre-test, but had converged to option (D) (53% of answers) by the post-test. This convergence towards (D) was driven by the PbT group (67% of responses), where the TbP group was split between (A), (B) and (D). It is not entirely clear what drove this convergence, though the presence of an airtight membrane in (D) and a belief that an air gap is ultimately beneficial may have contributed. On the other hand, Q14 was not very difficult and most students, regardless of group, correctly answered (B).

There was only one question, Q1, with a fall in mean scores, so it is worth understanding why this occurred. The correct answer was (B), but in both tests the most frequent answer was (E) (frequency of $E_{pre}=12$ and $E_{post}=13$), with little difference between groups. Indeed, the number of students selecting the correct answer *fell* after participation from 8 to 6, driven primarily by the TbP group. There are two possible explanations for this: (1) the form of the question is confusing due to the use of "if any" and "not" and (2) thinking of concrete's high capacitance as a source of "capacitive insulation". Indeed, one could argue that the overall effect of dampened indoor temperatures resulting from the use of exposed concrete coupled with effective night-ventilation might seem to many students as essentially performing the role of an "insulator". This is likely to have been reinforced by our focus on the *beneficial* effects of the interaction of mass and insulation in both the theory and practice elements (supported by our observations for Q11 and Q17 earlier), particularly in shelter construction for a hot climate, and less of a focus on the poor insulative properties of concrete in normal buildings. This may have further reinforced the highly positive *cultural* association of thermal mass in vernacular construction as the only source of passive thermal regulation. These considerations suggest that student understanding is mediated by not only the form of the questions and small shifts in emphasis during delivery, but also the *locale* of instruction and its cultural underpinnings.

Discussion and conclusion

The overall aim of our paper is to discuss the use of carefully designed pedagogical experiments that lead to evidence-based decisions in curriculum design, particularly when the design can be interpreted as a binary choice. Our focus is to address the question of whether it is better to begin with theory and build up to practice, i.e. the use of "capstone" projects, or vice versa—the use of a "foundation stone". We consider the various issues surrounding the design of an experiment to answer such a question, starting with ensuring fairness, consistency and comparability.

We lay out the key steps needed to design and execute such experiments and discuss several subtle issues to do with the definition of terms and attributes. We suggest that for an experiment of this nature to be deemed "successful" it would be ideal to use formal assessment as a metric and for the resultant difference between groups means to be of the order of 5%. An analysis of effect sizes and a simple power analysis suggests per-group sample sizes of between 51 and 128 students to observe a large to medium effect for such a difference in means, at the 0.05 confidence level. Crucially, we suggest that this difference in means should be judged with respect to a baseline measure of knowledge in order to determine the true impact of differential sequencing, particularly in the case of smaller samples. We discuss the strengths and weaknesses of synchronous and non-synchronous experimental formats, with the former providing more experimental control but with the risk of smaller sample sizes, compared to the latter.

We undertake an experiment of the synchronous form as an end-to-end demonstration of our ideas. This includes the development of novel learning materials specifically designed to test our hypothesis in the field of architectural engineering, using approaches common in the discipline (Dutson et al. 1997). While we recognise that practice is many-valued and complex (Hager et al. 2012; Reich et al. 2015), we use the well-known, relatively narrow,

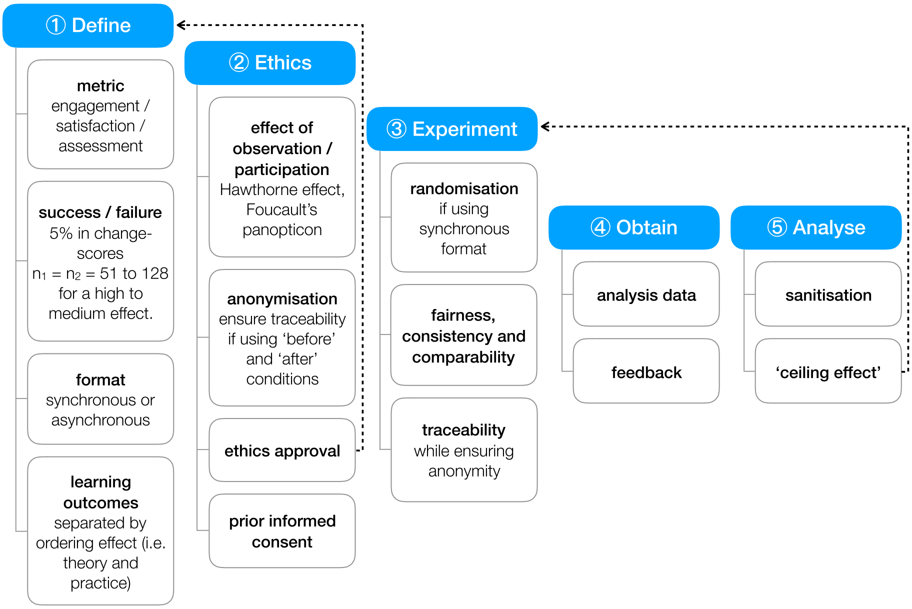

① Define
metric
engagement / satisfaction / assessment
success / failure
5% in change-scores
n₁ = n₂ = 51 to 128 for a high to medium effect.
format
synchronous or asynchronous
learning outcomes
separated by ordering effect (i.e. theory and practice)
② Ethics
effect of observation / participation
Hawthorne effect, Foucault's panopticon
anonymisation
ensure traceability if using 'before' and 'after' conditions
ethics approval
prior informed consent
③ Experiment
randomisation
if using synchronous format
fairness, consistency and comparability
traceability
while ensuring anonymity
④ Obtain
analysis data
feedback
⑤ Analyse
sanitisation
'ceiling effect'

Fig. 4 Suggested revisions (dotted lines) to the framework proposed in Fig. 1, as possible pathways for further iterative development

definition of practice as the "process of applying knowledge towards the solution of a real-world problem" (Dutson et al. 1997). To this end, we chose a topic of international importance—the design of shelters for refugees—as ideal for testing our hypothesis due to its constrained nature. The experiment is conducted with students in Jordan, a major host of Syrian refugees, which is one of the largest refugee populations in the world at present. The first, and simplest, outcome of this is that a "foundation stone" approach, i.e., practice *before* theory, is possible. A second outcome is the possibility that the framework presented in Fig. 1, is susceptible of modification and iteration based on experience. For example, a reviewer has correctly suggested that the ethical review process in Step 02 might lead to a re-evaluation of Step 01 "Define". Similarly, although we did not observe a limitation emerging from the ceiling effect, preliminary testing in other work may well lead to the need for a re-evaluation of the testing process. Both these are now indicated in Fig. 4.

Although sample sizes were small ($n_{TbP}=11$ and $n_{PbT}=9$), suggesting the need for a larger experiment, the difference in change–scores between groups was 6% (in favour of the "Practice before Theory" group), consistent with our suggestion of a 5% threshold. Mean change–score standard deviation was 14%, which is on the higher end of the spectrum of standard deviations we discuss in "Defining: success or failure" section. Although this is unsurprising for such small samples, they provide a useful first estimate for future studies when undertaking power analysis. The experiment highlights the importance of using change scores rather than raw summative results, since the latter might show significance when none truly exists, especially for small sample sizes.

Deeper analysis of attainment in individual questions provided a more "textured" understanding of the conceptual gains made by the students as a result of participation in the learning activities. Disregarding grouping demonstrates that overall attainment improved by 27 percentage points in all but three questions out of seventeen. Questions centred around the core focus of the learning activities, i.e. the joint use of insulation and thermal mass showed extremely large increases (62% and 71%), with little difference between groups, suggesting successful engagement within both groups. The two questions demonstrating no change in scores involved concepts that the students either did not understand at all, or found relatively simple. The only question that resulted in a drop in score may have been due to a combination of difficult wording, shifted tutor emphasis and cultural effects.

Finally, we demonstrate that student feedback needs to be considered in addition to attainment, as the choice will become less clear if the two are in conflict. Hence, our approach can be considered as robust, and designed to ensure there is strong evidence in favour of a binary choice decision in curriculum design.

Acknowledgements This research was partially supported by EPSRC funded Healthy Housing for the Displaced project (Grant ref EP/P029175/1) and an ERASMUS training grant (Grant 2017-1-UK01-KA107-035699). Neither EPSRC nor the ERASMUS programme were involved in the design of the study. For others wishing to either replicate or extend this experiment, a copy of the test, the SSST and the anonymised score data for the students is included in the data repository at http://doi.bath.ac.uk/. We would also like to thank the two anonymous reviewers for their helpful comments in improving the manuscript, and also the associate editor for their enthusiasm and support of the paper.

References

Banios, E. W. (1992). Teaching engineering practices. In *Frontiers in education conference* (pp. 161–168). IEEE.

Dutson, A. J., Todd, R. H., Magleby, S. P., & Sorensen, C. D. (1997). A review of literature on teaching engineering design through project-oriented capstone courses. *Journal of Engineering Education, 86,* 17–28.

Fenwick, T. (2012). Matterings of knowing and doing: Sociomaterial approaches to understanding practice. In P. Hager, A. Lee, & A. Reich (Eds.), *Practice, learning and change* (pp. 67–83). Berlin: Springer.

Fenwick, T., & Edwards, R. (2010). *Actor-network theory in education*. Abingdon: Routledge.

Fenwick, T., Edwards, R., & Sawchuk, P. (2015). *Emerging approaches to educational research: Tracing the socio-material*. Abingdon: Routledge.

Fosas, D., Albadra, D., Natarajan, S., & Coley, D. (2017). Overheating and health risks in refugee shelters: Assessment and relative importance of design parameters. In *Proceedings of the 33rd PLEA international conference: design to thrive. PLEA* (pp. 3746–3753).

Fosas, D., Albadra, D., Natarajan, S., & Coley, D. A. (2018). Refugee housing through cyclic design. *Architectural Science Review, 61,* 327–337.

Gander, R., Salt, J., & Huff, G. (1994). An electrical engineering design course sequence using a top-down design methodology. *IEEE Transactions on Education, 37,* 30–35.

Gherardi, S. (2009). Introduction: The critical power of the practice lens. *Management Learning, 40*(2), 115–128.

Hager, P., Lee, A., & Reich, A. (2012). Problematising practice, reconceptualising learning and imagining change. In P. Hager, A. Lee, & A. Reich (Eds.), *Practice, learning and change* (pp. 1–14). Berlin: Springer.

Hill, G. A. (2017). The 'tutorless' design studio: A radical experiment in blended learning. *Journal of Problem Based Learning in Higher Education, 5,* 111–125.

Kearsley, G., & Shneiderman, B. (1998). Engagement theory: A framework for technology-based teaching and learning. *Educational Technology, 38,* 20–23.

Liebman, J. C. (1989). Designing the design engineer. *Journal of Professional Issues in Engineering, 115,* 261–270.

Reich, A., Rooney, D., Gardner, A., Willey, K., Boud, D., & Fitzgerald, T. (2015). Engineers' professional learning: A practice-theory perspective. *European Journal of Engineering Education, 40,* 366–379.

Schatzki, T. R. (2012). A primer on practices: Theory and research. In D. Sheehan & J. Higgs (Eds.), *Practice-based education* (pp. 13–26). Leiden: Brill Sense.

Törley, G. (2014). Algorithm visualization in teaching practice. *Acta Didactica Napocensia, 7,* 1–17.

UNHCR (2016). *Shelter design catalogue*. Technical report.

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