

Interactions between defining, explaining and classifying: The case of increasing and decreasing sequences

Lara Alcock · Adrian Simpson

Received: date / Accepted: date

Abstract This paper describes a study in which we investigated relationships between defining mathematical concepts — increasing and decreasing infinite sequences — explaining their meanings, and classifying consistently with formal definitions. We explored the effect of defining, explaining, or studying a definition on subsequent classification, and the effect of classifying on subsequent explaining and defining. We report that 1) student-generated definitions and explanations were highly variable in content and quality; 2) explicitly considering the meaning of the concept facilitated subsequent classification, and giving a personal definition or explanation had a greater effect than studying a given definition; 3) classifying before defining or explaining resulted in significantly poorer definitions and explanations. We discuss the implications of these results for the teaching of abstract pure mathematics, relating our discussion to existing work on the concept image/concept definition distinction and on working with examples.

Keywords Sequences · Definitions · Examples · Real Analysis · Classification

1 Introduction

Definitions are central to contemporary formal mathematics because, in order to develop deductive arguments and to communicate clearly, mathematicians need to agree upon precise meanings for mathematical concepts. Historically these meanings have been debated (Lakatos, 1976), but contemporary mathematical norms dictate that once a definition is agreed, it is essentially indisputable; revolutions in meaning remain possible (Brown, 1998) but, at least for learners of mature theories, membership of mathematical categories should be entirely determined by whether objects satisfy definitional properties. This means that mathematical definitions are stipulative rather than

L. Alcock

Lara Alcock, Mathematics Education Centre, Loughborough University, Loughborough, LE11 3TU, UK
E-mail: L.J.Alcock@lboro.ac.uk

A. Simpson

(✉) School of Education, Durham University, Leazes Lane, Durham, DH1 1TA, UK.
E-mail: adrian.simpson@durham.ac.uk

lexical (Vinner, 1991), and that their associated categories are ‘classical’: that is, they are not formally treated as having internal structure (for example, objects do not formally have ‘greater’ or ‘lesser’ degrees of membership), and they have clearly-defined boundaries across which idiosyncratic classifications are not allowed.

In contrast, everyday natural categories are not specified by stipulative definitions, and there is evidence to suggest that human beings decide about membership of natural categories not by checking against properties but by judging on the basis of similarity to prototypes or to previously encountered exemplars (Kruschke, 2005). In natural categories, therefore, some objects are commonly considered better or worse examples, boundaries may be ill-defined, and classification can be a somewhat personal decision — idiosyncratic classifications, especially of ‘boundary examples’, are to be expected.

This difference between mathematical and natural categories is commonly understood to cause student errors, because students’ concept images are unlikely to conform precisely to the corresponding defined concepts (Tall & Vinner, 1981; Vinner, 1991). At a global level, habituation to natural categories means that students might not think to invoke definitions when trying to classify objects or construct mathematical arguments (e.g. Alcock & Simpson, 2004; Moore, 1994). At a local level, it means that students’ ideas about category membership might be skewed towards common types of examples, so that their answers are inconsistent with defined concepts (Fujita, 2012; Vinner, 1991) and even, in some cases, inconsistent with one another (Alcock & Simpson, 2011). Advanced mathematics thus requires that students learn to classify and argue using standard concept definitions (Vinner, 1991) and to modify pre-existing concept images so that they are consistent with defined concepts.

For educators, this raises the question of how teaching might help. There is general skepticism about the extent to which the stereotypical undergraduate ‘definition-theorem-proof’ pedagogy (cf. Weber, 2004) achieves what is required. In such pedagogy, the lecturer or textbook introduces a definition, perhaps gives an example or two, then proceeds as though membership of the defined category is unproblematic or simply a technical issue. Evidence suggests that this leaves room for ambiguity about the way in which definitions are supposed to determine categories. Bergé (2008), for instance, found that three consecutive course texts in one institution formally defined the same concept in logically equivalent but distinct ways; Raman (2004) contrasted textbooks that used only informal statements, those that stated formal definitions but make little use of them, and those that stated formal definitions and used them extensively. Both authors argued that such ambiguity makes it difficult for students to understand how definitions should be interpreted and used.

In response to these concerns, some researchers in mathematics education have developed and studied interventions to support students in understanding the role and use of definitions. Some have designed computer environments to promote understanding of logically complex statements (Dubinsky, Elterman, & Gong, 1988); others have argued that traditional pedagogy gives students little opportunity to develop expert-like knowledge of a wide range of examples (Moore, 1994) and suggested that more extensive, structured experience with examples might be beneficial (Fukawa-Connelly & Newton, 2014; Watson & Mason, 2005). There is also considerable interest in involving students in the defining process by asking them to work collaboratively to construct or reconstruct definitions (e.g. Moore-Russo, Conner, & Rugg, 2011; Swinyard, 2011; Zandieh & Rasmussen, 2010). Researchers working in this last way have described their participants’ reasoning in terms of Lakatosian principles and constructs (e.g. Larsen & Zandieh, 2008; Ouvrier-Buffet, 2011), and provided insights into the criteria that students and teachers employ when discussing the relative

merits of different possible definitions (e.g. Dawkins, 2014; Oehrtman, Swinyard, & Martin, 2014; Zaslavsky & Shir, 2005).

In this paper we also take an approach that deepens our understanding of defining as a mathematical activity, but that does so by investigating short-term interactions with specific concepts. We designed this approach after reconsidering the methods used in studies that have explored the concept image/concept definition distinction. In our view, this extensive literature has provided useful information on typical concept images for a range of important concepts, but has missed an opportunity to use systematic variation in task sequencing to explore the interactions between defining and working with examples. In the next section, we explain how.

2 Task Sequencing in Definition-Based Studies

Research on specific concept images (e.g. Biza, Christou, & Zachariades, 2008; Fujita, 2012; Moore, 1994; Vinner, 1991) has tended to draw on evidence from two types of task:

- Classifying tasks and reasoning tasks (such as proof construction) in which participants should invoke a definition but are not explicitly told to do so;
- Defining tasks in which participants are asked to state the meaning of a concept, either as a formal definition or as a personal explanation.

The order in which these tasks have been used has depended on the researchers' aims.

In some studies, researchers have presented the classifying or reasoning task first and the defining task second, because they wished to find out whether participants would spontaneously invoke and correctly apply a definition; they postponed the defining task because they did not want to prompt this behaviour, but they did want to check whether participants knew the definition and so, in principle, could have used it. For example, Vinner and Dreyfus (1989) asked first year college and junior high school students to classify graphs and written descriptions (including atypical examples) according to whether or not they represented functions. They subsequently asked the students, 'What is a function, in your opinion?', and found that even students whose responses to this latter task were consistent with the formal definition classified in ways that did not conform to it. With younger pupils, Fujita (2012) posed a number of questions that generally began with a task such as classifying quadrilaterals into families and ended with a defining task ('What is a parallelogram? Describe it in words.'). He too found that students appeared to work from prototypical examples when classifying, even when their descriptions were consistent with the formal definitions.

Other studies have used the opposite order, giving a defining task first and a classifying or reasoning task second so as to study the extent to which participants can use a definition when prompted. Inglis and Simpson (2008), for instance, presented both mathematics students and members of a general, well-educated population with a relatively complex disjunctive definition, then required them to use it in a classifying task. They found that while mathematics students appeared to classify better in this circumstance, there remained evidence that many relied upon exemplars. Similarly, Biza et al. (2008) first asked 12th grade students to explain what is meant by the term 'tangent line', then set various tasks related to tangents in a calculus context, including classifying lines according to whether or not they were tangents to given functions (again these tasks included atypical examples). They described students' behaviour as relying on concept images, which appeared to be heavily influenced by earlier experience with tangents in Euclidean geometry.

Other researchers have used different task combinations. Alcock and Simpson (2011), for instance, began by asking new undergraduates to complete a classifying task for which the concepts,

increasing and *decreasing* for sequences, had not been formally introduced. They then presented the standard definitions and asked participants to repeat the classifying task with these definitions clearly visible. They reported that performance in the second round was still not entirely consistent with the definitions, and moreover that for some participants it was not possible to associate responses with any single, consistently applied criterion.

We make two methodological observations about these studies.

First, in each one, the main focus was on participants' responses to the classifying or reasoning task: the aim was either to explore concept images for their own sake, or to explore reasoning after a prompt to think about a definition. Thus the defining task typically played a subordinate role: in define-second studies, it was usually reported only as a check; in define-first studies, it acted as a check and/or a prompt. Reports of these studies therefore tend to downplay the fact that providing a definition itself involves significant cognition; they do not typically examine the nature of the responses beyond considering how well these match the relevant definition.

Moreover, defining tasks have varied considerably: some studies provided participants with a formal definition (Alcock & Simpson, 2011; Inglis & Simpson, 2008), some asked for a formal definition (Heinze & Kwak, 2002), some asked for something more akin to an explanation ('try to explain, in simple words...' Biza et al., 2008), while some were still more flexible ('define or explain as you wish' Bingolbali & Monaghan, 2008). We suggest that responses might be influenced by the prompt type: we might expect, for instance, that a request to explain would result in more idiosyncratic and imprecise responses than a request to define. The present study explores this question, using three basic prompt types: read a definition, give a definition, and explain the meaning.

Our second methodological observation is that these studies have typically used the same task order for all participants, meaning that they do not provide information on whether and how experience of the first task affects responses to the second. We suggest that this constitutes a missed opportunity to gather important information: being asked to define or explain a concept might well influence approaches to a subsequent classifying or reasoning task, and conversely recent reasoning attempts or exposure to a range of examples — especially atypical ones — might influence subsequent attempts to define or explain. Both theoretically and pedagogically, such potential interactions are of interest: it would be useful to better understand what we should expect as learners make the transition to advanced mathematics so that we could make more informed choices about task ordering. We thus also manipulated the order of defining tasks and a corresponding classifying task, so that our study overall addressed the following research questions.

1. How do students respond when asked to define or explain the meaning of *increasing* and *decreasing* for sequences?
2. What effect does classifying have on subsequent defining or explaining?
3. What effect does defining or explaining have on subsequent classifying?

We used the sequence concepts *increasing* and *decreasing* for three reasons. First, Alcock and Simpson (2011) found that in relation to these concepts, students classified sequences in a wide variety of ways. Second, their formal definitions are relatively straightforward compared with many definitions at this level (each has only one quantifier and one relation), so poor classifications are less likely to result from failing to understand the definitions (Dubinsky & Yiparaki, 2000). Third, because their everyday meanings are accessible, a participant who is asked to explain, or one who cannot remember a standard definition, should still be able to give a sensible response.

3 Method

Our first research question suggests an exploratory stance to this part of the work and, as with previous literature detailed above, we designed a study in which participants were asked to define or explain a concept. We then used qualitative analysis to uncover patterns within those responses, and we describe and illustrate both common and uncommon response types in Section 4.1. Research questions 2 and 3, however, are about facilitation or interference effects: they explore whether experience of one task results in systematic differences in subsequent responses to another. We thus develop our quantitative analysis from that experimental design, the results from which appear in Sections 4.3 and 4.4.

3.1 Participants

The participants for this study were 132 first year undergraduate students at a high-ranking UK university. All were studying for degrees in either mathematics or natural sciences; both degree programmes involved a significant formal mathematics component taught in standard definition-theorem-proof style. This approach remains the norm for large lectures in this department, and indeed across UK undergraduate mathematics teaching.

Data were collected in week 8 of the first ten-week term, after the students had experienced some formal mathematics and while all were taking a real analysis module. This means that prior to the study, the students had met a number of formal definitions concerning sequences of real numbers. In particular, four weeks before the study they were introduced to the notions of increasing and decreasing; they were given a formal definition (in the same form as that used in the study) and introduced to a small number of applications, including a demonstration that a constant sequence is formally classified as both increasing and decreasing.

3.2 Tasks

The participants were randomly assigned to one of five seminar groups for an associated module. At the start of a seminar each participant was given a booklet containing two tasks (a classifying task and a defining task) in set order as described below. Participants were asked to complete the task booklet, undertaking the tasks in the order given and avoiding returning to a task or section once they had completed it.

The classifying task was the same for all participants so we describe this first. The task required participants to classify fifteen sequences as increasing, decreasing, both or neither; Table 1 gives the list of sequences and the normatively correct responses. The choice of sequences is partly based on those used by Alcock and Simpson (2011) and each booklet had the list of sequences in random order.

The defining task varied between groups. Participants in one group were provided with formal definitions of increasing and decreasing and asked to study these; students in the other groups were asked either to define these terms or to explain their meanings. Groups, tasks, and the orders in which the tasks were completed are summarized in Table 2. Note that we have reproduced only the wording of the defining tasks for *increasing*; the wording for *decreasing* tasks was otherwise identical and, as there were very few differences in character of the responses for *increasing* and *decreasing*, the analysis focuses only on the former.

Table 1 Classifying Task Sequences

Sequence	Correct classification
$(0, 1, 0, 1, 0, 1, 0, 1, \dots)$	Neither
$(1, 4, 9, 16, 25, 36, 49, 64, \dots)$	Increasing
$(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}, \dots)$	Decreasing
$(1, -1, 2, -2, 3, -3, 4, -4, \dots)$	Neither
$(3, 3, 3, 3, 3, 3, 3, \dots)$	Both/Neither (*)
$(1, 3, 2, 4, 3, 5, 4, 6, \dots)$	Neither
$(6, 6, 7, 7, 8, 8, 9, 9, \dots)$	Increasing/Neither (*)
$(0, 1, 0, 2, 0, 3, 0, 4, \dots)$	Neither
$(10\frac{1}{2}, 10\frac{3}{4}, 10\frac{7}{8}, 10\frac{15}{16}, 10\frac{31}{32}, 10\frac{63}{64}, 10\frac{127}{128}, 10\frac{255}{256}, \dots)$	Increasing
$(-2, -4, -6, -8, -10, -12, -14, -16, \dots)$	Decreasing
(a_n) where $a_n = n^2 + 1$	Increasing
(a_n) where $a_n = 3 + (-1)^n$	Neither
(a_n) where $a_n = 5 - \frac{1}{2^n}$	Increasing
The sequence shown in the graph below:	
	Neither
The sequence shown in the graph below:	
	Increasing

3.3 Coding

There is one complexity associated with the context chosen for the research: there are two standard definitions for *increasing* and *decreasing*, where the difference depends on whether the inequality between consecutive terms is strict or not. Two sequences in the classifying task (highlighted with asterisks in Table 1) would be categorized differently according to these definitions, but were

important to include given previous research which showed interesting responses to these types of behaviour (e.g. Roh, 2008). In the quantitative analysis, the responses were scored (out of 15) against both the strict and non-strict definitions and awarded the higher of the two marks. This means that we took the most conservative approach to avoid introducing bias, which could only decrease any effects which might be uncovered.

Table 2 Groups and task ordering.

Group		First task	Second task
Given-Classify ($N = 27$)	(GC)	“The following is the definition of what it means for a sequence (a_n) to be increasing: A sequence (a_n) is increasing if $\forall n \in \mathbb{N} \ a_{n+1} \geq a_n$ ”	Classifying task
Define-Classify ($N = 25$)	(DC)	“Define carefully what it means for a sequence (a_n) to be increasing: A sequence (a_n) is increasing if . . .”	Classifying task
Explain-Classify ($N = 26$)	(EC)	“What does it mean for a sequence to be increasing?”	Classifying task
Classify-Define ($N = 29$)	(CD)	Classifying task	“Define carefully what it means for a sequence (a_n) to be increasing: A sequence (a_n) is increasing if . . .”
Classify-Explain ($N = 25$)	(CE)	Classifying task	“What does it mean for a sequence to be increasing?”

For the four groups asked to define or explain (total $N = 105$), each participant was awarded a define/explain score of 2, 1 or 0:

- 2: if the relationship between consecutive sequence terms was sensibly represented and properly quantified;
- 1: if the relationship between consecutive sequence terms was sensibly represented but quantification was omitted or incorrect;
- 0: if neither the relationship between consecutive terms nor the quantification was adequately captured.

Each response was scored independently by two experienced mathematics lecturers. There was a high level of agreement on the score awarded (weighted $\kappa = 0.787, p < 0.001$) with perfect agreement in 81% of cases. In the remaining cases, the two raters agreed a final define/explain score. They tended to be lenient with ambiguous quantification (awarding a score of 2) but strict with a total absence of quantification or quantification over an incorrect set (awarding a score of 1, provided the inequality relationship was clear). Illustrative scored responses are given in Figure 1 which also indicates the group from which the response originated. Detailed information on variation in the responses is given in the Section 4.

Score	Examples
2	$x_n > x_{n-1} \quad \forall n \in \mathbb{N}_0$ (EC) $\frac{a_{n+1}}{a_n} > 1 \quad \forall n \in \mathbb{N}$ $a_{n+1} > a_n$ (DC)
1	$a_{n+1} > a_n$ (CD) $a_{(n+1)}$ is greater than a_n . (CD)
0	<i>its getting bigger</i> (CD) <i>if $f'(x) > 0$ ($a_n^{x \in (a,b)}$) then increasing</i> (CE)

Fig. 1 Illustrations of scores for define/explain tasks. Letters in parentheses indicate groups as listed in Table 2.

4 Analysis and Results

4.1 Responses to define/explain tasks: Qualitative analysis

In this section we give a qualitative overview of the define/explain responses from all relevant groups.

Responses to the define/explain tasks varied in a number of ways, and two features accounted for much of this variation: the forms of the inequality relationship ($a_{n+1} > a_n$) and the quantified statement about the indexing set ($\forall n \in \mathbb{N}$). The inequality relationship was sometimes expressed using an inequality symbol (e.g. ' $a_{n+1} > a_n$ ') and sometimes in words (e.g. ' a_{n+1} is bigger than a_n ' or 'the term is larger than the previous one'). We categorise these below as 'symbolic relation' and 'verbal relation' respectively. The indexing set was either completely specified (e.g. 'for all n in \mathbb{N} '), incomplete (e.g. 'for all n ', 'increasing n ', 'each' or ' n in \mathbb{N} '), or missing altogether. Where quantification was incomplete, participants commonly delimited the range of the relationship using different phrases or with chains of inequalities, quantified across terms of the sequence rather than across the indexing set ('for all a_n ' instead of 'for all n '), or quantified over the incorrect set (\mathbb{Z} or \mathbb{R} instead of \mathbb{N}). These differences are illustrated in the examples in Figure 2.

$a_{n+1} > a_n \quad \forall n \in \mathbb{N}$ (DC)	The term is a_{n+1} is greater than a_n for all terms $a_1 < a_{n_1} < a_{n_2} < a_{n_3} \dots < a_n$ (CE)	$a_{n+1} \geq a_n$ (CD)
A sequence x_n is increasing iff $x_{n+1} \geq x_n \quad \forall n \in \mathbb{N}$. (EC)	a_{n+1} is bigger than $a_n \quad \forall n \geq 1$ (CD)	$a_n \leq a_{n+1}$ (DC)
(a) Complete quantification	$a_n < a_{n+1}$ for every a_n $n \in \mathbb{R}$ (CD)	(c) Missing quantification
(b) Incomplete or incorrect quantification		

Fig. 2 Illustrations of different uses of quantification. Letters in parentheses indicate groups as listed in Table 2.

A total of 89 responses could be classified on these two dimensions, as summarized in Table 3.

Table 3 Responses by relation and quantifier type.

	Quantifier Type		
	Complete	Incomplete	Missing
Symbolic relation	11	17	17
Verbal relation	0	34	10

There were 16 remaining responses which were too idiosyncratic to be classified in this way. These are particularly interesting from the point of view of qualitative analysis, because they did not capture the concepts in conventional form, but did suggest that participants had some awareness of the meaning of increasing and were attempting to capture this in a concise form. Some appeared to invoke a meaning from another context (such as differentiable functions or scatterplots); others used an expression that did not indicate any context (as in “goes up without going down”). Examples of both types are illustrated in Figure 3, which gives a flavour of the breadth of variation in these responses. Overall, in addressing our first research question, the study revealed a surprisingly wide range of responses to the define-explain task and highlighted the differences between individual sense-making and formal definitions.

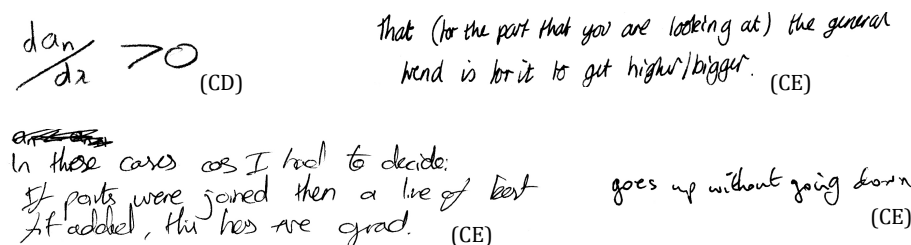


Fig. 3 Illustrative idiosyncratic responses.

4.2 Responses to define/explain tasks: Between-group differences

The complexity of the data above makes it challenging to discern systematic relationships between the tasks the students were asked to undertake and the wide response variation. We thus focussed first whether there was an effect of group on relation type (symbolic/verbal) and quantification type (complete/incomplete/missing). Count data for this question are shown in Table 4.

A Fisher’s exact test did not show a significant relationship between group and quantification type ($p = 0.230$, Fisher’s exact test). However, there was significant difference between relation type (symbolic/verbal) across the groups ($p < 0.001$, Fisher’s exact test). An examination of standardised residuals showed that the only significant contributor to this difference was the classify-explain group

Table 4 Relation and quantifier type by group.

Group	Relation Type		Quantifier Type			Unclassifiable
	Symbolic	Verbal	Complete	Incomplete	Missing	
DC	17	6	5	8	10	2
EC	13	10	3	16	4	3
CD	13	13	2	16	8	3
CE	2	15	1	11	5	8

($z = +2.28$ verbal and $z = -2.25$ symbolic). That is, after classifying, students were more likely to give a verbal response to the request to explain the meaning.

This provides a first indication of interference between the tasks: classifying before being asked to explain led to different forms of explanation. The remaining results sections explore other ways in which the tasks may interfere, using the define/explain and classify scores to address research questions 2 and 3.

4.3 Effects of classifying on defining/explaining

For our second research question, we had expected those asked to define to attain significantly higher scores than those asked to explain the meanings of the concepts, but a Mann-Whitney test showed that there was no significant difference between the define/explain scores of those asked to explain and those asked to define ($U = 1192, z = -1.306, p = 0.192$).

A similar test did, however, show a significant difference between those who were asked to define/explain before classifying and those asked to do so after classifying. The groups asked to define/explain first scored significantly higher than those asked to define/explain second ($U = 984, z = -2.781, p < 0.01$) — summary information is shown in Table 5. We calculated the effect size using a point estimation of Vargha and Delaney’s A , giving an unbiased estimate of stochastic superiority of the define/explain-first group over the define/explain-second group of $\hat{A}_{\text{first, second}} = 0.643$, which would be considered the equivalent of a medium effect size with parametric statistics (Vargha & Delaney, 2000). Equivalently, the average student in the define/explain-second group would stand around the 69th percentile in the define/explain-first group. Thus, the experience of classifying the sequences before being asked to define or explain seems to have interfered with students’ abilities to give a clear statement in response to the define/explain task.

Table 5 Mean define/explain score for each group.

Group	Mean score	σ
DC	1.31	0.471
EC	1.31	0.679
CD	1.11	0.737
CE	0.72	0.678

For our second research question, the results provide evidence about the interactions between defining tasks and classifying tasks, suggesting that recent exposure to examples may disrupt students' abilities to give clear statements of meaning. This might be the result of an availability bias (Pollard, 1982): those who have just completed a classifying task involving atypical examples are likely to have those examples in mind when trying to capture the essence of a concept, and this might disrupt their response to a define/explain task. This might apply particularly to the classify-explain group, who could have been trying to articulate post-hoc justifications for idiosyncratic classifications: the requirement to explain could have allowed these participants permission to describe what they had done rather than to define the concept, and thus led to the significantly larger number of verbal responses.

4.4 Effects of defining/explaining on classifying

There was a significant but weak positive correlation between the define/explain score and the classifying score ($\rho = 0.339, p < 0.001$). That is, those who gave better definitions/explanations tended to classify in a way more consistent with the formal definitions. For the participants as a whole ($N = 132$) the mean classification score was 11.76 ($\sigma = 2.64$). Table 6 gives the means and standard deviations of the scores for each group, and for two larger groupings based on whether the participants classified first or defined/explained first.

Table 6 Between-groups comparison of classification score.

Group	N	Mean	σ	N	Mean	σ	Grouping
GC	27	11.81	2.74	79	12.53	2.57	define/explain first
DC	25	12.38	2.61				
EC	26	13.42	2.14				
CD	29	10.82	2.13	53	10.60	2.34	classify first
CE	25	10.36	2.58				

A Mann-Whitney test indicated that classifying scores were significantly greater for those asked to classify second than for those asked to classify first ($U = 1143.5, z = -4.449, p < 0.001$). We calculated the effect size using a point estimation of Vargha and Delaney's A . This gives the unbiased estimate of stochastic superiority of the classify-second group over the classify-first group of $\hat{A}_{\text{first, second}} = 0.726$ which would be considered the equivalent of a large effect size with parametric statistics (Vargha & Delaney, 2000): the average student in the classify-first group would stand around the 80th percentile in the classify-second group. Note, however, that the scores of all three classify-second groups left some room for improvement, even though this study involved highly qualified students working on a simple concept with which they had some experience. Like other studies that have presented definition tasks first, this suggests that learners must effortfully override concept-image based reasoning in order to behave mathematically in relation to defined concepts.

Further insight can be obtained by examining these results in more detail, because the form of the define/explain task also had a significant influence on the score: a Kruskal-Wallis test showed the mean ranks of scores were different among groups ($\chi^2 = 25.9, df = 4, p < 0.001$). Since we were examining the students' classifications, we conflated the two groups who classified first (as

there could be no influence on their classification caused by the second task). A set of post-hoc Mann-Whitney tests on each pair of resulting groups, with Holm sequential Bonferroni adjustment for significance (with $\alpha = 0.05$), showed that differences between given-classify and define-classify groups and between the explain-classify and define-classify groups were not significant. However, the explain-classify and the define-classify groups scored significantly higher than those who classified first (EC: $U = 244, z = -4.674, p < 0.001$; DC: $U = 398, z = -3.058, p = 0.002$). In addition, the explain-classify group scored significantly higher than the given-classify group ($U = 210, z = -2.567, p = 0.010$).

Thus those who classified second classified better regardless of whether their first task required defining or explaining. In fact, those who were asked to explain the meanings of the terms significantly outperformed those who were asked to study the provided formal definition. This is somewhat surprising: one might think that a request to explain would lead to lower quality and more idiosyncratic interpretations of the concepts, and would thus result in poorer classification.

Overall we observe that no group had perfect classification scores, indicating that even when specifically prompted, highly qualified students working with simple definitions do not all behave in a formal mathematical way in relation to defined concepts. But focusing first on formulating meanings for the concepts facilitated classification, and this held whether the task involved reading a given definition, providing a definition, or explaining the meanings of the concepts. Indeed, the requirement to study a given formal definition is apparently a weaker support for correct classification than the requirement to explain, at least for the students in our sample: those who explained attained higher classification scores than those who studied a provided definition. At first this might appear counterintuitive: the given definition was precise and stated in a form that participants had previously seen, so one might expect that asking for an explanation would invite more idiosyncratic classifications. Our result to the contrary might indicate that the requirement to explain cues a thoughtful attempt to give a statement that is personally meaningful, while the invitation to read a definition might not engage the student in such deep consideration. Overall, it may be that the key factor is the extent to which the task prompts students to focus on fixing a meaning for the concept by writing down their definition or explanation.

5 Discussion

5.1 Implications for research

Understanding defined concepts involves two linked demands: one global and one local. The global demand is that students learn to treat definitions as stipulative criteria, and to invoke and use them appropriately in response to reasoning tasks. The local demand is that students learn to work with each particular definition and conform their concept image to it. We know that this process is difficult as a whole, because previous research has repeatedly shown that students new to advanced mathematics do not respond in line with definitions on classifying and reasoning tasks. This is also the case for the students in our study. But we do not know whether the problem is primarily global – that students do not treat definitions as stipulative so they do not invoke them when appropriate – or whether it is primarily local – that they do attempt to invoke definitions but fail to do so successfully in some cases, or that they invoke definitions successfully but do not apply them correctly to particular examples.

Discussions using the concept image/concept definition distinction often take the problem to be primarily global (e.g Fujita, 2012; Moore, 1994; Tall & Vinner, 1981). They describe the main

difficulty as that of making the transition from a world in which categories are natural and definitions are descriptive (and may be safely ignored once a good concept image exists) to a world in which categories are defined and definitions are stipulative. However, it is not obvious that this is the case. The abstract idea of a stipulative definition is not very complicated: small-scale studies have demonstrated that teachers and successful students employ sensible criteria when stating and discussing the relative merits of different possible definitions (e.g. Dawkins, 2014; Zaslavsky & Shir, 2005); they seem to expect that a definition should apply universally to every instance of a given concept. So perhaps the problem is primarily local, and the main difficulty is that of applying any particular definition, especially when this conflicts with existing imagery.

We do not claim that our study fully resolves this question, as either position is broadly consistent with our results. Our participants were highly qualified and the relevant definitions have simple logical structures, so it seems plausible that the participants should not have had difficulty with the technical details and are more likely to have failed by not using a definitional approach; on the other hand, we deliberately included atypical examples, so local conflicts were likely to arise and unlikely to be resolved perfectly by all participants. We claim, instead, that by varying task sequences we have provided evidence that both global and local problems exert competing influences on students' reasoning. In our study, classifying first led to a larger proportion of participants giving verbal rather than symbolic responses, and to definitions/explanations that were less consistent with the standard formal definition. This indicates that local effects are operating: students are swayed by recent experience with atypical examples. Also, however, defining or explaining first led to classifications more consistent with a normatively correct interpretation of the terms: apparently students can discipline themselves to treat a definition as stipulative, even if they do not always do so perfectly.

We recognise, of course, that our between-subjects design does not provide information on the learning trajectories of individuals. Perhaps most students learn quickly that they should treat definitions as stipulative, but continue for some time to make errors with regard to specific concepts; perhaps most gradually bring their understandings into line for specific concepts without really being aware of a global shift. Perhaps there are large individual differences, with some students behaving in a normatively correct way from early in their experience of advanced mathematics and others never learning to treat definitions as stipulative. Further research would be needed to untangle these possibilities, and we suggest that a variety of methodological approaches would be necessary. For example, it might also be instructive to relate student performance on tasks like ours to performance in standard assessments, to provide evidence about the extent to which correct interpretation of definitions predicts performance in a broader array of tasks in advanced mathematics. But we do think that our findings both open up this issue and sound an important note of caution for future studies on the relationship between concept image and concept definition: task order matters.

5.2 Implications for teaching and learning

Stereotypical university mathematics teaching gives primacy to formal definitions (Bergqvist, 2007; Weber, 2004), while much mathematics education literature emphasises enhancing example spaces so that they more accurately coincide with the extensions of the corresponding definitions (Fukawa-Connelly & Newton, 2014; Goldenberg & Mason, 2008). We note that suggestions about example spaces have often been based on theoretical characterisations of expert understanding, rather than on empirical research into ways in which exposure to a definition influences judgments about ex-

amples or ways in which exposure to examples influences the production of definitions. The present study highlights the potentially intricate co-dependencies between classifying and defining.

Like other studies, ours shows that definitions are often not well remembered and not spontaneously invoked, and that prompts to consider definitions do not result in ideal classifications. Unlike other studies, it provides direct evidence that although exposure to atypical examples may be desirable in terms of developing an expert-like concept image, it might, at least temporarily, disrupt an ability to focus on definitions. This needs to be considered carefully in instructional sequences that involve extensive experience with examples. Perhaps such sequences should be interleaved with tasks that demand explicit attention to definitions, such as prompts to check an answer against a formal definition and, if appropriate, to change it. For lecturers designing instruction for students new to advanced mathematics, further research that systematically varies prompt types and orders could be illuminating.

Indeed, there is much scope for further exploration of both task sequencing and task types. While our study used a range of define/explain tasks (read a formal definition, state a definition, explain the meaning of a term), we gave only one form of reasoning task: a set of examples to classify. It might be useful to investigate whether co-dependencies between task responses are also present in situations where students are asked to think about meanings of terms before or after completing tasks other than classifying. There has, for instance, been some debate on the value of example generation for supporting reasoning (Dahlberg & Housman, 1997; Iannone, Inglis, Mejía-Ramos, Simpson, & Weber, 2011; Watson & Mason, 2005). While spontaneous example generation could well differ in its effects from non-spontaneous example generation (Sandefur, Mason, Stylianides, & Watson, 2013; Yopp, 2014), it would be interesting to compare ways in which defining tasks and example generation tasks influence further reasoning, and to investigate the impact of different defining tasks on example generation. Further studies exploring interactions among factors present in a variety of pedagogical situations could both aid our understanding and help us design tasks and instructional sequences to assist students in making an effective transition to advanced mathematics.

References

- Alcock, L., & Simpson, A. (2004). Convergence of sequences and series: Interactions between visual reasoning and the learner's beliefs about their own role. *Educational Studies in Mathematics*, *57*(1), 1–32.
- Alcock, L., & Simpson, A. (2011). Classification and concept consistency. *Canadian Journal of Science, Mathematics and Technology Education*, *11*(2), 91–106.
- Bergé, A. (2008). The completeness property of the set of real numbers in the transition from calculus to analysis. *Educational Studies in Mathematics*, *67*(3), 217–235.
- Bergqvist, E. (2007). Types of reasoning required in university exams in mathematics. *The Journal of Mathematical Behavior*, *26*(4), 348–370.
- Bingolbali, E., & Monaghan, J. (2008). Concept image revisited. *Educational Studies in Mathematics*, *68*(1), 19–35.
- Biza, I., Christou, C., & Zachariades, T. (2008). Student perspectives on the relationship between a curve and its tangent in the transition from euclidean geometry to analysis. *Research in Mathematics Education*, *10*(1), 53–70.
- Brown, J. R. (1998). What is a definition? *Foundations of Science*, *3*(1), 111–132.
- Dahlberg, R. P., & Housman, D. L. (1997). Facilitating learning events through example generation. *Educational Studies in Mathematics*, *33*(3), 283–299.

- Dawkins, P. C. (2014). How students interpret and enact inquiry-oriented defining practices in undergraduate real analysis. *The Journal of Mathematical Behavior*, *33*, 88–105.
- Dubinsky, E., Elterman, F., & Gong, C. (1988). The student's construction of quantification. *For the Learning of Mathematics*, *8*(2), 44–51.
- Dubinsky, E., & Yiparaki, O. (2000). On student understanding of AE and EA quantification. *Research in Collegiate Mathematics IV*, 239–289.
- Fujita, T. (2012). Learners' level of understanding of the inclusion relations of quadrilaterals and prototype phenomenon. *The Journal of Mathematical Behavior*, *31*(1), 60–72.
- Fukawa-Connelly, T. P., & Newton, C. (2014). Analyzing the teaching of advanced mathematics courses via the enacted example space. *Educational Studies in Mathematics*, *87*(3), 323–349.
- Goldenberg, P., & Mason, J. (2008). Shedding light on and with example spaces. *Educational Studies in Mathematics*, *69*(2), 183–194.
- Heinze, A., & Kwak, J. Y. (2002). Informal prerequisites for informal proofs. *Zentralblatt für Didaktik der Mathematik*, *34*(1), 9–16.
- Iannone, P., Inglis, M., Mejía-Ramos, J. P., Simpson, A., & Weber, K. (2011). Does generating examples aid proof production? *Educational studies in Mathematics*, *77*(1), 1–14.
- Inglis, M., & Simpson, A. (2008). Reasoning from features or exemplars. In O. Figueras, J. Cortina, S. Alatorre, T. Rojana, & A. Sepúlveda (Eds.), *Proceedings of the 32nd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 217–224). Morelia, Michoacán, Mexico.
- Kruschke, J. K. (2005). Category learning. In K. Lamberts & R. Goldstone (Eds.), *The handbook of cognition* (pp. 183–201). London: Sage.
- Lakatos, I. (1976). *Proofs and refutations: The logic of mathematical discovery*. Cambridge University Press.
- Larsen, S., & Zandieh, M. (2008). Proofs and refutations in the undergraduate mathematics classroom. *Educational Studies in Mathematics*, *67*(3), 205–216.
- Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics*, *27*(3), 249–266.
- Moore-Russo, D., Conner, A., & Rugg, K. I. (2011). Can slope be negative in 3-space? studying concept image of slope through collective definition construction. *Educational Studies in Mathematics*, *76*(1), 3–21.
- Oehrtman, M., Swinyard, C., & Martin, J. (2014). Problems and solutions in students' reinvention of a definition for sequence convergence. *The Journal of Mathematical Behavior*, *33*, 131–148.
- Ouvrier-Buffet, C. (2011). A mathematical experience involving defining processes: in-action definitions and zero-definitions. *Educational Studies in Mathematics*, *76*(2), 165–182.
- Pollard, P. (1982). Human reasoning: Some possible effects of availability. *Cognition*, *12*(1), 65–96.
- Raman, M. (2004). Epistemological messages conveyed by three high-school and college mathematics textbooks. *The Journal of Mathematical Behavior*, *23*(4), 389–404.
- Roh, K. H. (2008). Students' images and their understanding of definitions of the limit of a sequence. *Educational Studies in Mathematics*, *69*(3), 217–233.
- Sandefur, J., Mason, J., Stylianides, G., & Watson, A. (2013). Generating and using examples in the proving process. *Educational Studies in Mathematics*, *83*(3), 323–340.
- Swinyard, C. (2011). Reinventing the formal definition of limit: The case of Amy and Mike. *The Journal of Mathematical Behavior*, *30*(2), 93–114.
- Tall, D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, *12*(2), 151–169.

- Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25(2), 101–132.
- Vinner, S. (1991). The role of definitions in the teaching and learning of mathematics. In D. Tall (Ed.), *Advanced mathematical thinking* (pp. 65–81). Dordrecht: Kluwer Academic Publishers.
- Vinner, S., & Dreyfus, T. (1989). Images and definitions for the concept of function. *Journal for Research in Mathematics Education*, 356–366.
- Watson, A., & Mason, J. (2005). *Mathematics as a constructive activity: The role of learner generated examples*. Mahwah, NJ: Erlbaum.
- Weber, K. (2004). Traditional instruction in advanced mathematics courses: A case study of one professor's lectures and proofs in an introductory real analysis course. *The Journal of Mathematical Behavior*, 23(2), 115–133.
- Yopp, D. A. (2014). Undergraduates' use of examples in online discussions. *The Journal of Mathematical Behavior*, 33, 180–191.
- Zandieh, M., & Rasmussen, C. (2010). Defining as a mathematical activity: A framework for characterizing progress from informal to more formal ways of reasoning. *The Journal of Mathematical Behavior*, 29(2), 57–75.
- Zaslavsky, O., & Shir, K. (2005). Students' conceptions of a mathematical definition. *Journal for Research in Mathematics Education*, 317–346.