Semantic contamination and mathematical proof: Can a non-proof prove? [1]

Juan Pablo Mejía-Ramos [a] Matthew Inglis [b]

[a] *Graduate School of Education, Rutgers University*

[b] *Mathematics Education Centre, Loughborough University*

Abstract

The way words are used in natural language can influence how the same words are understood by students in formal educational contexts. Here we argue that this so-called semantic contamination effect plays a role in determining how students engage with mathematical proof, a fundamental aspect of learning mathematics. Analyses of responses to argument evaluation tasks suggest that students may hold two different and contradictory conceptions of proof: one related to conviction, and one to validity. We demonstrate that these two conceptions can be preferentially elicited by making apparently irrelevant linguistic changes to task instructions. After analyzing the occurrence of "proof" and "prove" in natural language, we report two experiments that suggest that the noun form privileges evaluations related to validity, and that the verb form privileges evaluations related to conviction. In short, we show that (what is judged to be) a non-proof can sometimes (be judged to) prove.

Key words:
language, mathematics, proof, reasoning, semantic contamination

Argumentation and proof are widely accepted as being central to mathematics (Heinze and Reiss, 2007; Hilbert et al., 2008). Not surprisingly then, educators generally agree that argumentation and proof should be incorporated into mathematics learning and instruction at all levels (Hanna, 2007; Schoenfeld, 1994). However, many studies have shown that students find engaging with proof difficult, regardless of whether such engagement takes the form of evaluating given proofs or constructing novel proofs (Coe and Ruthven, 1994; Fischbein, 1982; Harel and Sowder, 1998; Knuth, 2002; Selden and Selden, 2003). There is a long tradition in the literature of trying to account for these difficulties by looking at individual students' conceptions of proof, and finding

[1] We are grateful to James Adelman, David Pimm, Keith Weber, and an anonymous referee for their helpful comments on this work.

mismatches between these and the agreed standards in the discipline (Bell, 1976; Healy and Hoyles, 2000; Segal, 1999). Our goal in this paper is to build on this long research tradition, but also to attempt to connect it to a rather different and under-researched area: the linguistics of mathematical proof. We first discuss some theoretical background issues associated with both these areas.

1 Semantic contamination.

The influence of language on the learning has been widely recognized in the education literature. It is now accepted that the way linguistic structures are used in natural language (by which we mean language from day-to-day life) can differ from how the same linguistic structures are used in formal contexts and, in particular, in mathematics. When analyzing this phenomena, Halliday (1975) used the term "register" to refer to "a set of meanings that is appropriate to a particular function of language, together with the words and structures which express those meanings" (p. 65). Halliday suggested that the mathematical register is relatively unusual for a technical register, in that when naming new phenomena, it often redefines simple words from natural language rather than coining novel technical terminology. This can give rise to what Pimm (1987) called *semantic contamination*: where the meaning or usage of a term from natural language influences how the term is understood by a learner in the mathematical register.

Several examples of semantic contamination have been discussed in the mathematics education literature. In the context of advanced mathematics for example, Monaghan (1991) found that the natural language meaning of words and phrases associated with the limit concept ("tends to", "approaches", "converges", etc.) can impact upon students' concept images of the formal limit concept in calculus and analysis classes. Similarly, Tall and Vinner (1981) suggested that colloquial meanings of the term "continuity" influence how students engage with the formal mathematical concept. Difficulties which arise from such issues will need to be overcome if the learner is to successfully engage with the mathematical register (Pimm, 1987; Schleppegrell, 2007).

In this paper we explore whether students' understanding of mathematical proof is influenced by semantic contamination. In other words, we ask does the way that words associated with proof are used in natural language influence how they are understood in the mathematical register?

To establish semantic contamination three steps are required. The first is to investigate the way in which the to-be-analyzed concept is referred to in natural language. In the case of limit discussed by Monaghan (1991), for example, the natural language use and meanings of the terms “limit”, “approaches”, “tends to” and so on were investigated. The second step is to empirically examine how these concepts are understood by students in mathematical contexts: Monaghan conducted a survey which interrogated students’ interpretations of these terms. The third and final step is to argue that the results from the analyses conducted in the first two stages correspond sufficiently to suggest that natural language referents to the concept influences students’ understanding of the concept in mathematical contexts.

We begin this three step process by analyzing the contexts in which the concept of proof appears in natural language.

2 Proof in natural language.

In the English language there are two main linguistic methods of referring to the concept of proof: it can appear as a noun (“proof”) or as a verb (“prove”). To investigate the way that these noun and verb forms are used in natural language, we searched for instances of the noun form (“proof” or “proofs”) and the verb form (“prove” and “proves”) in the British National Corpus (BNC) World Edition (Burnard, 2000). The BNC is a comprehensive collection of 100 million words of spoken and written English, designed to represent a cross-section of current English usage. Analyses of large scale corpora, such as the BNC, are widely used by researchers interested in the usage patterns of various linguistic features. In particular, such techniques have been highly productive at comparing the use of words in different registers (Stubbs, 2004).

Our aim was to determine the frequency of the verb and noun referents to proof in specialist language (i.e. language normally associated with a specific formal context or topic, such as education, business, legal, medical, etc.) and informal day-to-day language (i.e. language which could be found spoken on popular radio or in informal conversations, and so on). In Halliday’s (1975) sense, we aimed to compare the occurrences of the verb and noun referents to proof in typical formal and informal registers.

The spoken component of the BNC consists of approximately 10 million words split into two sections: impromptu conversational spoken English recorded from the day-to-day life of 124 representative volunteers, and spoken English recorded from timetabled events in various different contexts: educational (e.g. university lectures), business (e.g. trade union talks or business meetings), public/institutional (e.g. parliamentary proceedings) and leisure (e.g. sports commentaries). We formed a Specialist Language category by grouping the educational, business and public/institutional context domains, and an Informal Language category by grouping the conversational component and the

Table 1
Frequency of noun and verb forms of proof in the spoken component of the BNC, by the language-type of the source.

	Noun	Verb
Specialist Language	121 (54%)	105 (46%)
Informal Language	51 (35%)	95 (65%)

leisure context domain.

Table 1 shows the frequency of the noun and verb referents in the spoken component of the BNC, separated by the language-type of the source document (specialist or informal, defined as described above). [2] The noun form was found significantly more often in specialist language (54% of occurrences of proof in specialist language were of the noun form) compared to informal language (where 35% of occurrences were of the noun form), $\chi^2(1) = 12.355, p < .001$. [3] Importantly, this disproportionate occurrence of the noun form of proof in specialist language does not merely reflect an overall overabundance of nouns in formal contexts. In fact, nouns tend to be slightly less common than average in specialist language (Hudson, 1994).

These data indicate that in spoken natural language the noun form tends to be most often found in specialist, formal registers. In contrast, the verb form is most often found in non-specialist informal registers. Our goal in this paper is to investigate whether these different patterns of use influence how learners engage with the terms "proof" and "prove" in the mathematical register. In other words, is the way that students engage with mathematical proof influenced by semantic contamination?

Before reporting our empirical work regarding this issue, we first situate the paper within existing literature on proof. Specifically, we review evidence that students (at both school and university levels) have two simultaneous and contradictory conceptions of proof which can cause conflict in the learning of mathematics.

[2] This word frequency analysis included several unusual meanings of "proof" and "prove" (for example "proof" can refer to the aeration of dough by a raising agent before baking). Because semantic contamination refers to how the ways in which a word is used in natural language influences how it is understood in the mathematical register, we did not attempt to remove any such atypical uses (which could have an impact upon the nature of any semantic contamination into the mathematical register) from the sample.
[3] An equivalent analysis on the written component of the BNC revealed a similar pattern of results.

3 Two aspects of proof: conviction and validity.

The importance of knowing about students' conceptions of proof was pointed out by Bell (1976) as part of his investigation into school-children's mathematical explanations. Bell suggested that proof serves multiple purposes in the discipline of mathematics, and that a student is unlikely to produce the type of argument required by his or her teacher unless they agree on which purpose the to-be-produced argument must satisfy. Such considerations may well account for the robust finding that students at all levels often make favorable judgements about empirical arguments (the checking of a number of examples) during argument evaluation tasks (e.g. Harel and Sowder 1998; Knuth 2002). Although such arguments are often seen as being very convincing, they are typically perceived as being invalid. There is now clear evidence that students can independently consider both these purposes of proof during evaluation tasks.

As different researchers have used subtly different terminology to refer to similar phenomena, we will make explicit the terms used in this paper (which are consistent with those used by Segal 1999). An argument is seen as *convincing* if it, or some translation of it, raises the reader's level of belief in its conclusion above some given threshold. In contrast, an argument is seen as *valid* if the reader believes that it meets the agreed criteria set down by the mathematical community; typically this would involve accurate logic and being based on appropriate premises (the prototypical example would be a deductive argument). We recognize that conviction and validity may be linked for a particular argument. For example an argument may be convincing by virtue of its validity. Similarly an argument may be convincing because the reader sees that some transformation of it would be valid, even if it itself is not. However, arguments do exist which are usually viewed by mathematicians as being convincing and invalid, where no obvious translation of the argument is valid; see, for example empirical arguments related to the Riemann Hypothesis (Borwein et al. 2008, chpt. 4), or visual arguments which are far removed from their formal counterparts (Nelsen, 1997).

In a study of the proof conceptions of university students, Segal (1999) investigated first-year undergraduates conceptions of proof. She operationalized the notion of conviction by asking students to read a series of mathematical arguments and select for each one either "this convinces me of the result", or "this does not convince me of the result". Further instructions emphasized the nature of the judgement Segal was asking her participants to make: "you are asked to decide whether *you personally* find the argument convincing" (p. 198, emphasis in the original). The notion of validity was operationalized in an analogous fashion: students read the instruction "*a proof is an argument by which one persuades one's enemies* (as opposed to one's friends, or one-self)" (p. 199, emphasis in the original), and then were asked to select either "this proves the result" or "this does not prove the result" for each argument.

Segal (1999) found that students' responses to each of these questions depended on the type of argument they were assessing. When evaluating empirical arguments, many students seemed to make their judgements about conviction and validity using different criteria. Midway through their first year of undergraduate studies, over half of her sample considered each empirical argument in the study to be convincing but invalid. In contrast students tended to reach identical judgements about how convincing and valid they found deductive arguments. One account of these findings is that students used separate criteria for conviction and validity for both empirical and deductive arguments, but that these judgements simply coincided for the deductive argument in the study. That is to say that while empirical arguments rated highly for conviction and lowly for validity, the students saw deductive arguments as both convincing and valid.

A similar study at school level was conducted by Healy and Hoyles (2000). In a large-scale investigation of 14- and 15-year-old children's argumentation and proof behavior, Healy and Hoyles again found that students simultaneously hold two different conceptions of proof: those which they themselves would adopt (related to conviction), and those they believe would receive the best mark (related to validity). For example, students for the most part selected narrative and empirical arguments when asked to pick arguments closest to what their own approach would be; but when asked to pick those arguments which would get the best mark, they would predominantly choose deductive algebraic arguments. Healy and Hoyles drew the same conclusion as Segal; that their participants simultaneously held two different conceptions of proof, each of which could be elicited during argument evaluation tasks using carefully designed questions.

Given the theoretical importance of Segal's (1999) and Healy and Hoyles's (2000) studies, it is worth noting that there has been surprisingly little research which has investigated what factors influence whether students adopt a conviction conception of proof or a validity conception of proof. Segal found that asking students about personal conviction privileges conviction and asking students about persuading enemies privileges validity; but if Bell (1976) was correct, then many of the problems students have with mathematical proof stem from confusing the validity and conviction conceptions. But it is unclear what factors influence which conception students adopt when they are asked about proof with a relatively neutral question that does not clearly privilege one conception or the other. In order to develop effective pedagogy it would be particularly valuable to determine whether the conception students adopt is influenced by trait variables (i.e. relatively stable individual characteristics and beliefs of the student), by state variables (i.e. properties of the environment that the student finds themselves in), or by some combination of the two.

Given the two main linguistic ways of representing the concept of proof ("proof" and "prove"), and the different conceptions of proof evoked during argument evaluation tasks (conviction and validity), it is natural to ask whether there is any link between these two dualities. In other words, does semantic contamination from natural language influence how the words "proof" and "prove" are understood in the mathematical register?

The BNC analysis suggests a plausible hypothesis for the nature of any possible semantic contamination effect. If such an effect did occur, we might expect the noun form ("proof") to more often evoke the (more formal) validity conception of proof (because of its disproportionate occurrence in formal technical registers), and the verb form ("prove") to more often evoke the (more personal) conviction conception of proof (because of its disproportionate occurrence in everyday informal registers). The main goal of the experiments reported in the remainder of this paper was to investigate the possibility of such a relationship.

4 Experiment 1: Can a non-proof prove?

The primary goal of Experiment 1 was to determine whether the conviction and validity conceptions of proof are related to the two different linguistic referents of the concept of proof. Specifically, we asked: does the noun referent to proof privilege the validity conception, and does the verb referent privilege the conviction conception?

To explore this issue, following Segal (1999) and Healy and Hoyles (2000), we decided to investigate students' responses to argument evaluation tasks. Developing the ability of students to evaluate arguments is a key goal of mathematics education. Indeed, Selden and Selden (2003) suggested that evaluating purported proofs for validity is the primary method by which many mathematicians learn new mathematics. Weber (2008) argued that the activity is also important for both students and teachers. In fact developing students' skills at evaluating mathematics arguments is an explicit goal of most curricula: the NCTM (2000) suggested that teachers should instigate a community of inquiry by helping students to critique their classmates' arguments (p. 346). In short, evaluating arguments is a important mathematical activity that students are expected to engage with. Consequently it serves as a suitable place for testing the hypotheses derived in the preceding analysis.

4.1 Procedure and participants.

We used the internet to gather data in order to maximize our sample size. Although such an approach does create potential problems that need to be carefully considered (Reips, 2002), well-designed web-based experiments have been found to produce data that is consistent with lab-based methods (Gosling et al., 2004; Krantz and Dalal, 2000). Following the practice adopted by Inglis and Mejia-Ramos (2009), we recorded the IP address of each participant, together with the time they submitted their responses. Under the assumption that each IP address was associated with a unique individual, these data were used to screen for possible cases of multiple submission.

Participants were 220 volunteers studying mathematics at one of twenty highly-ranked US universities (selected from the USNews.com list of "top mathematics programs"). Each participant was recruited via an email sent by their departmental secretary which explained the purpose of the study, and asked them to visit the experimental website should they wish to participate. Before starting the experiment participants were asked to declare that they were mathematics students and that they had not previously participated in the study.

4.2 Materials.

The study followed a between-subjects design to minimize any influence of participants' perceptions about the purpose of the study. After having made their declaration participants were randomly assigned to participate in either the noun condition or the verb condition. They were then presented with a claim-argument pair and asked either "Is the argument a proof of the claim?" (noun condition) or "Does the argument prove the claim?" (verb condition). Responses were recorded via a straightforward yes/no tick box form. Once they had clicked submit, a second claim-argument pair was presented and the same question asked.

Two different claim-argument pairs were used. The experimental claim-argument pair was a visual argument that purportedly justified the series

$$\sum_{n=1}^{\infty} \frac{1}{2^{2n}} = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots = \frac{1}{3}.$$

We chose the visual argument carefully. Although visual arguments in mathematics are epistemologically controversial (Brown, 2008; Giaquinto, 2007), most commonly they are seen as being mathematically invalid (Folina, 1999);

Table 2
Frequency of responses to the visual argument about series convergence, by condition.

	Yes	No
Noun	53 (46%)	63 (54%)
Verb	62 (60%)	42 (40%)

and such critical views are often held by students in more extreme forms than by mathematicians (Inglis and Mejia-Ramos, 2009). In contrast visual arguments are often seen to be highly convincing (Knuth, 2002). The visual argument we chose is typical in this respect.[4] Given this, we would expect a different range of responses to this argument from those participants who adopt the conviction conception of proof (from whom we would expect proportionately more 'yes' responses) to those who adopt the validity conception (from whom we would expect proportionately more 'no' responses).

In addition we used a filler task (an indirect argument in favor of the claim "if n^2 is divisible by 3, then n is divisible by 3"; adapted from an argument used by Selden and Selden 2003) in order to increase the length of the study (experience suggests that very short internet-based experiments have greater problems with data fidelity; Reips 2002). The two claim-argument pairs are given in full, as they appeared in the study, in the Appendix. The order in which the two claim-argument pairs were displayed was randomized for each participant.

4.3 Results.

Participants' responses to the visual series argument are shown in Table 2. Participants' conditions did influence their responses, $\chi^2(1) = 4.262, p = .039$. A small majority of participants in the noun condition responded 'no' (54%), whereas the situation was reversed in the verb condition, where 60% responded 'yes'.

For completeness, we also analyzed the data from the filler task, in which participants were asked about the indirect argument. These data are shown in Table 3. Surprisingly, there was also a significant between-conditions difference on this argument, $\chi^2(1) = 4.502, p = .034$. Large majorities of participants in

[4] Furthermore, the argument is convincing *not* because it can easily be translated into a valid argument. A typical formal proof of this statement requires a careful examination of the series's sequence of partial sums, and selection of an appropriate n such that the n-th partial sum is arbitrarily close to $\frac{1}{3}$. The image does not suggest a successful way of making this selection.

Table 3
Frequency of responses to the indirect argument about divisibility by 3, by condition.

	Yes	No
Noun	96 (83%)	20 (17%)
Verb	96 (92%)	8 (8%)

both conditions responded positively to the argument, but significantly fewer participants did so in the noun condition (83%) than in the verb condition (92%).

4.4 Discussion.

Participants' responses to the visual argument in Experiment 1 seemed to confirm the hypothesis generated from the BNC analysis. Those in the noun condition most often said that they did not think that the visual argument was a proof, as we would expect if they were biased towards evaluating the validity of the argument. In contrast those in the verb condition most often said that they thought that the argument did prove the claim, as we would expect if they were biased towards evaluating how convincing the argument was.

The data from the filler indirect argument were rather surprising, for two reasons. First, the argument was adapted from one used by Selden and Selden (2003) in their study of eight undergraduates' proof validation behavior. Selden and Selden found that the ability of their participants to determine that this argument was valid was extremely limited, and initially no better than chance. In contrast we found that a large majority (87%) of our participants rated the proof positively. Second, if the link between noun/verb and validity/conviction that we have proposed were the case, we would not have expected any difference in ratings between the conditions, as we would have expected that the argument would be seen as both valid and convincing (i.e. unlike the visual argument it would receive positive responses regardless of which conception of proof participants adopted). In fact we found a significant difference between the two conditions. We believe that there are two reasonable ways of accounting for this finding.

One possibility is that, despite the evidence from the BNC analysis, in fact the distinction between the verb and noun forms of proof is not related to validity and conviction at all, but merely to the level of evidence required for a positive evaluation. For instance, suppose participants were only judging the validity of the argument, but that they did so with varying degrees of certainty.[5] Suppose further that, rather than privileging validity, the noun form of the question instead led participants to require a comparatively higher threshold of certainty about the validity of the argument before they would assign it proof status. In contrast, perhaps the verb form required a comparatively lower threshold of certainty for proof status. That is to say that the reason for the between-conditions difference that we found may have been that the participants were sufficiently certain about the argument's validity for it to pass the lower threshold engendered by the verb condition, but that they were not sufficiently certain for it to pass the higher threshold engendered by the noun condition.

A second possibility is that the indirect nature of the argument caused the differences between the two conditions. The argument, purportedly in support of the claim "if n^2 is divisible by 3, then n is divisible by 3", actually established the claim's contrapositive ("if n is not divisible by 3, then n^2 is not divisible by 3") before asserting the equivalence of the claim and its contrapositive (see Appendix). Perhaps a subset of participants in the noun condition answered "no" to the question "is the argument a proof of the claim?" because they saw the argument as actually being a proof of the contrapositive of the claim (cf., Goetting 1995). Such an interpretation would not apply to participants in the verb condition (who answered the question "does the argument prove the claim?") as, even if they had seen the argument as a proof of the contrapositive of the claim rather than the claim, this would be sufficient to prove the claim.

To explore these competing accounts we conducted a second experiment.

5 Experiment 2.

The main aims of Experiment 2 were to (i) replicate the primary result of Experiment 1 with a second argument; and (ii) to determine whether the result from the indirect argument in Experiment 1 would also be found with a direct argument of similar difficulty level.

5.1 Procedure, participants and materials.

The procedure was identical to that used in Experiment 1, with the exception of the content of two arguments. In place of the original visual argument an

[5] A similar threshold account based on conviction (or indeed some linear combination of conviction and validity) could be constructed. Such accounts are interchangeable in terms of the predictions tested in Experiment 2.

Table 4
Frequency of responses to the visual argument for Young's Inequality, by condition.

	Yes	No
Noun	40 (31%)	88 (69%)
Verb	50 (44%)	63 (56%)

visual argument in support of Young's Inequality was used. In place of the indirect argument we used a direct formal argument which demonstrates that the product of two diagonal matrices is diagonal (based on the argument used by Segal 1999). Both arguments are given in full in the Appendix.

We chose the direct proof carefully. If the 'certainty threshold' account of the data from the indirect argument was correct we would expect similar results with the new direct proof, as the nature of the argument would not affect the hypothesized different threshold levels associated with the noun and verb conditions. If, in contrast, the indirect nature of the indirect argument caused the between-conditions effect, we would expect it to be abolished on the direct argument in Experiment 2.

Participants were 241 volunteers studying mathematics at one of twenty highly-ranked US universities (again selected from the USNews.com list of "top mathematics programs", but different to those which participated in Experiment 1). As before, participants were contacted by their departmental secretary via email and asked to visit the experimental website should they wish to participate.

5.2 Results.

Participants' responses to the visual argument for Young's Inequality are shown in Table 4. Participants' conditions again influenced their responses: 31% of participants in the noun condition responded 'yes', compared to 44% in the verb condition, $\chi^2(1) = 4.333, p = .037$.

Responses to the direct argument are shown in Table 5. Large majorities of participants in both conditions (91% and 90% for noun and verb conditions respectively) ranked the argument favorably. The between-conditions difference

Table 5
Frequency of responses to the direct argument about diagonal matrices, by condition.

	Yes	No
Noun	116 (91%)	12 (9%)
Verb	102 (90%)	11 (10%)

did not approach significance, $\chi^2(1) = 0.009, p = .924$. [6]

5.3 *Discussion.*

Participants' responses to the visual argument for Young's Inequality were broadly similar to those for the visual argument in Experiment 1. Participants in the noun condition were less likely to rate the argument positively than those in the verb condition. Exactly this pattern of results would be expected if, as suggested by the BNC analysis, the verb referent to proof privileged evaluations related to conviction and the noun referent privileged evaluations related to validity.

The data from the direct argument are consistent with the suggestion that it was the indirect nature of the filler-task in Experiment 1 which caused the between-conditions difference. That is to say that a minority of participants were of the view that the indirect argument in Experiment 1 was a proof of the contrapositive of the claim, and not a proof of the claim itself. It is hard to reconcile the competing 'certainty threshold' account with the data from the direct argument in Experiment 2. If the noun condition made participants require more certainty than the verb condition, we would have expected to find a between-conditions difference, with higher numbers of participants rating the argument positively in the verb condition compared to the noun condition. In fact, we found a very small and non-significant trend in the opposite direction.

[6] That this comparison had similar power characteristics and distance from ceiling as the indirect proof comparison in Experiment 1, gives us increased confidence that a Type II statistical error has been avoided. Furthermore, our results are consistent with Segal's (1999) study where the same argument was given similar conviction and validity ratings by around 80% of participants. Recall that Segal used extremely explicit task instructions to ensure participants were making conviction and validity judgements respectively.

6 General Discussion.

6.1 Summary of main findings.

Our main goals in this paper were to determine whether the way learners engage with mathematical proof is affected by semantic contamination from natural language into the mathematical register. To accomplish this we explored Segal's (1999) and Healy and Hoyles's (2000) suggestion that some students hold two simultaneous and contradictory conceptions of mathematical proof: conviction and validity.

To establish that semantic contamination influences student understanding of a given concept, three steps are required. First one must analyze the natural language meanings and use of referents to the concept. Second one must investigate how the concept is understood by students in mathematical contexts. Finally one must argue that the natural language use of referents to the concept and students' understanding of the concept in mathematical contexts correspond sufficiently to suggest a causal relationship.

Our analysis of the word frequency of proof in natural language suggested that the noun referent to proof is more associated with specialist formal language than the verb referent, which disproportionately occurs in informal language. Consequently, under the hypothesis that semantic contamination is a factor in students' understanding of proof, we predicted that an argument evaluation task phrased using the noun form of proof ("proof") would elicit more responses which used the (more formal) validity conception of proof than the identical task phrased using the verb form ("prove"). In both Experiment 1 and Experiment 2, this was indeed the case for visual arguments. Participants in the verb conditions evaluated the arguments more favorably than those in the noun conditions: exactly what we would expect if they were disproportionately evaluating the argument in terms of conviction rather than validity.

The filler task used in Experiment 1 revealed a surprising aspect of indirect arguments. Although the large majority of participants in both conditions rated it positively, there was a significant between-conditions difference: with those in the noun condition more likely to give a negative rating. We have argued that this was because a small minority of participants seemed to believe that the argument was a proof of the contrapositive of the claim, and not a proof of the claim itself. This is not an implausible suggestion given the confusion some students have with indirect arguments (e.g., Goetting 1995). Responses to the direct argument in Experiment 2, where we found no between-conditions difference, were consistent with this account.

Although significant between-conditions differences were found for the visual arguments in Experiments 1 and 2, it is worth noting that the size of these effects was relatively small, $\phi = 0.139$ and $\phi = 0.134$ for Experiments 1 and 2 respectively. Although previous researchers have investigated semantic contamination in the context of mathematics, to our knowledge this is the study of the phenomena which has used a design that permits effect sizes to be calculated. Consequently, it is difficult to assess whether the size of the effects we found is comparable to the size of the semantic contamination effects in the areas of mathematics discussed by Monaghan (1991), Pimm (1987) and Tall and Vinner (1981).

An alternative way of characterizing these effect sizes is to note that substantial numbers of participants chose "yes" and "no" in both conditions. So, although the phrasing of the question—a state variable—did cause higher numbers of participants to adopt either the conviction or validity conception of proof, this factor on its own did not account for all the between-participants variability. Whether the remaining variability can be accounted for by further state variables related to the environment in which students took the task, or whether trait variables related to relatively stable beliefs and conceptions of individual students also play a part is unclear. Further research could profitably look at what other factors influence students' responses to argument evaluation tasks.

6.2 Implications.

Several researchers have noticed that the way terms are used in natural language can interfere with the way the same terms are understood within the mathematical register. This phenomena, referred to as semantic contamination by Pimm (1987), can cause problems for learners. In some cases students may fail to reason from formal concept definitions, and instead rely upon concept images influenced by natural language (Halliday, 1975; Monaghan, 1991; Tall and Vinner, 1981). To overcome this difficulty Rowland (1999) suggested that students must be helped to appropriate the technical language of the mathematical register into their own classroom discourse.

However, the case of mathematical proof seems to be somewhat different to the limit and continuity concepts discussed by Monaghan (1991) and Tall and Vinner (1981), as it does not have an agreed concept definition. In fact characterizing the nature of proof is a difficult and current topic in the philosophy literature. Some mathematicians have tried to remove considerations of conviction entirely, by characterizing proof as a sequence of formulae each of which either follows from previous formulae, or is an axiom (Hilbert, 1930); others have suggested that any convincing argument is a proof (Bundy et al., 2005). Others too, have argued that what is a proof depends on whether one is (or is being at the time) an applied mathematician, a pure mathematician or a computer scientist (Swinnerton-Dyer, 2005). Proof, then, would seem to be an unusual case of a technical term from the mathematical register with no clearly agreed concept definition.

One implication of this lack of an agreed meaning is that it seems very difficult to know whether or not we, as educators, should regard it as undesirable that students responded to the noun and verb forms of proof differently in the argument evaluation tasks reported in this paper. Without a clear reference from expert practice to guide our understanding of the mathematical register, it may be difficult to apply Rowland's (1999) suggestion and help our students appropriate the 'correct' meanings of the words "proof" and "prove".

Although it may be difficult to follow Rowland's (1999) suggestion in the context of proof, these findings do nevertheless have important implications for practice and theory. In particular Tall and Vinner's (1981) constructs of concept image and concept definition highlight to teachers the possibility that their students may be interpreting mathematical terms used in instruction differently to intended. For example, Monaghan (1991) showed that when lecturers refer to a sequence approaching a limit, students may believe that it must do so from only one side. In other words, an awareness of semantic contamination is important for real analysis lecturers as it highlights the possible communication breakdowns that may occur in their lectures. Similarly, the suggestion that semantic contamination occurs in the context of proof allows us, as teachers, to be aware of the conceptions of proof that we may be *unintentionally* evoking by our use of these terms in educational situations.

Along with practical implications, we believe that our findings have theoretical and methodological significance. Stylianides and Stylianides (2009) suggested that prospective student teachers sometimes construct empirical arguments in response to proof tasks, despite apparently regarding these very same self-produced arguments as invalid. Consequently, they argued that in order to gain a holistic picture of students' understanding of proof, researchers should use both evaluation and construction tasks. Our findings are consistent with the observation that it may be methodologically flawed to draw strong conclusions about an individuals' understanding of proof from responses to a single task. Indeed, we can go further, by saying that individuals may not even exhibit a stable single conception of proof across very similar argument evaluation tasks. As we randomly allocated our participants into conditions, our research design permits us to conclude that there is a causal relationship between the phrasing of argument evaluation task instructions and the types of responses that students give to them. That an apparently irrelevant linguistic change can influence participants' responses to standard argument evaluation tasks suggests that it may be unwise for educational theories to attribute stable mental constructs and beliefs to students in the domain of proof. In other words, it may be unwise to attempt to account for student behavior by considering trait variables only. We have shown that a seemingly innocuous state variable—whether the question uses a noun or a verb referent to proof—also influences student behavior. Future research could profitably investigate how state and trait variables interact in tasks related to mathematical proof.

Our empirical work here has concentrated on argument evaluation tasks: our participants were given arguments and asked to assess them. As discussed earlier in the paper, such evaluation activities are recognized as being central to providing students with an authentic mathematical experience and to developing their notions of argumentation and proof (NCTM, 2000; Selden and Selden, 2003; Weber, 2008). Nevertheless there are other important activities related to proof. In particular, in educational settings students are also often asked to construct valid mathematical arguments, and students often produce convincing but non-valid arguments in response to such tasks (Harel and Sowder, 1998; Recio and Godino, 2001). Does semantic contamination also influence student behavior in such settings? Would students be more likely to produce convincing but non-valid arguments if asked to prove a statement than if asked to write a proof of a statement? Given semantic contamination seems to influence behavior in argument evaluation tasks it seems reasonable to hypothesize that it would also influence behavior in argument construction tasks; but it will be for future research to test this hypothesis.

7 Conclusions.

We have demonstrated that whether an argument evaluation task uses the noun or verb referent to mathematical proof influences how students respond to it. We have accounted for this finding by proposing that semantic contamination between natural language and the mathematical register impacts upon how students engage with mathematical proof; that the conviction and validity conceptions of proof can be preferentially elicited by making an apparently innocuous linguistic change to task instructions. The noun referent seems to cue participants into evaluating arguments on validity grounds, whereas the verb referent cues participants into using conviction as a basis for their evaluations. In short, a non-proof can, in some cases, prove.

As well as increasing our understanding about the range of learning situations in which semantic contamination plays a role, these results have important implications for the significance of the observation that some students simultaneously hold two different conceptions of proof. They indicate one factor which influences which conception students adopt at any given time. Balacheff (2008) suggested that proof has several different meanings in the mathematics education literature, and pointed out that this could be a hindrance for research progress. The same point may be true for individual students and their mathematical progress. Understanding that different conceptions can be unintentionally elicited by the different linguistic choices made by teachers, lecturers and curriculum designers is surely a necessary prerequisite for helping our students successfully coordinate the various different aspects of mathematical proof.

References

Balacheff, N., 2008. The role of the researcher's epistemology in mathematics education: an essay on the case of proof. ZDM: The International Journal on Mathematics Education 40, 501–512.

Bell, A. W., 1976. A study of pupils' proof conceptions in mathematical situations. Educational Studies in Mathematics 7, 23–40.

Borwein, P., Choi, S., Rooney, B., Weirathmueller, A., 2008. The Riemann Hypothesis: A Resource for the Afficionado and Virtuoso Alike. Springer, New York.

Brown, J. R., 2008. Philosophy of Mathematics: A Contemporary Introduction to the World of Proofs and Pictures, 2nd Edition. Routledge, New York.

Bundy, A., Jamnik, M., Fugard, A., 2005. What is a proof? Philosophical Transactions of the Royal Society A 363, 2377–2391.

Burnard, L., 2000. Reference guide for the British National Corpus. Oxford University Computing Services, Oxford.

Coe, R., Ruthven, K., 1994. Proof practices and constructs of advanced mathematical students. British Educational Research Journal 20, 41–53.

Fischbein, E., 1982. Intuition and proof. For the Learning of Mathematics 3 (2), 9–18.

Folina, J., 1999. Pictures, proofs, and 'mathematical practice': Reply to James Robert Brown. British Journal for the Philosophy of Science 50, 425–429.

Giaquinto, M., 2007. Visual thinking in mathematics. Oxford University Press, Oxford.

Goetting, M., 1995. The college students' understanding of mathematical proof. Ph.D. thesis, University of Maryland.

Gosling, S. D., Vazire, S., Srivastava, S., John, O. P., 2004. Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. American Psychologist 59, 93–104.

Halliday, M. A. K., 1975. Some aspects of sociolinguistics. In: Jacobsen, E. (Ed.), Interactions between language and mathematical education. UNESCO, Paris, pp. 64–73.

Hanna, G., 2007. The ongoing value of proof. In: Boero, P. (Ed.), Theorems in school: From history, epistemology and cognition to classroom practice. Sense, Rotterdam, pp. 3–18.

Harel, G., Sowder, L., 1998. Students' proof schemes: Results from exploratory studies. In: Schoenfeld, A. H., Kaput, J., Dubinsky, E. (Eds.), Research in Collegiate Mathematics III. American Mathematical Society, Providence, RI, pp. 234–282.

Healy, L., Hoyles, C., 2000. A study of proof conceptions in algebra. Journal for Research in Mathematics Education 31, 396–428.

Heinze, A., Reiss, K., 2007. Reasoning and proof in the mathematics classroom reasoning and proof in the mathematics classroom. Analysis 27, 333–357.

Hilbert, D., 1930. Die grundlebung der elementahren zahlenlehre. Mathematische Annalen 104, 484–494.

Hilbert, T. S., Renkl, A., Kessler, S., Reiss, K., 2008. Learning to prove in geometry: Learning from heuristic examples and how it can be supported. Learning and Instruction 18, 54–65.

Hudson, R., 1994. About 37% of word-tokens are nouns. Language 70, 331–339.

Inglis, M., Mejia-Ramos, J. P., 2009. The effect of authority on the persuasiveness of mathematical arguments. Cognition and Instruction 27, 25–50.

Knuth, E., 2002. Secondary school mathematics teachers conceptions of proof. Journal for Research in Mathematics Education 33, 379–405.

Krantz, J. H., Dalal, R., 2000. Validity of web-based psychological research. In: Birnbaum, M. H. (Ed.), Psychological Experiments on the Internet. Academic Press, San Diego, pp. 35–60.

Monaghan, J., 1991. Problems with the language of limits. For the Learning of Mathematics 11 (3), 20–24.

NCTM, 2000. Principles and standards for school mathematics. NCTM, Reston, VA.

Nelsen, R. B., 1997. Proofs without Words: Exercises in Visual Thinking. MAA, Washington, DC.

Pimm, D., 1987. Speaking mathematically: Communication in mathematics classrooms. Routledge & Kegan Paul, London.

Recio, A., Godino, J., 2001. Institutional and personal meanings of mathematical proof. Educational Studies in Mathematics 48, 83–99.

Reips, U.-D., 2002. Standards for internet-based experimenting. Experimental Psychology 49, 243–256.

Rowland, T., 1999. The pragmatics of mathematics education: Vagueness and mathematical discourse. Falmer Press, London.

Schleppegrell, M. J., 2007. The linguistic challenges of mathematics teaching and learning: A research review. Reading & Writing Quarterly 23, 139–159.

Schoenfeld, A. H., 1994. What do we know about mathematics curricula? Journal of Mathematical Behavior 13, 55–80.

Segal, J., 1999. Learning about mathematical proof: conviction and validity. Journal of Mathematical Behavior 18, 191–210.

Selden, A., Selden, J., 2003. Validations of proofs considered as texts: can undergraduates tell whether an argument proves a theorem? Journal for Research in Mathematics Education 34, 4–36.

Stubbs, M., 2004. Language corpora. In: Davies, A., Elder, C. (Eds.), Hand-

book of Applied Linguistics. Blackwell, Oxford, pp. 106–132.
Stylianides, A. J., Stylianides, G. J., 2009. Proof constructions and evaluations. Educational Studies in Mathematics 72, 237–253.
Swinnerton-Dyer, P., 2005. The justification of mathematical statements. Philosophical Transactions of the Royal Society A 363, 2437–2447.
Tall, D. O., Vinner, S., 1981. Concept image and concept definition in mathematics with particular reference to limits and continuity. Educational Studies in Mathematics 12, 151–169.
Weber, K., 2008. How mathematicians determine if an argument is a valid proof. Journal for Research in Mathematics Education 39, 431–459.

A Arguments used in Experiment 1.

A.1 The visual series argument.

Claim. $\sum_{n=1}^{\infty} \frac{1}{2^{2n}} = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots = \frac{1}{3}$.

Argument. [Figure A.1].

A.2 The indirect argument.

Claim. For any positive integer n, if n^2 is divisible by 3, then n is divisible by 3.

Argument. We will show that the contrapositive of this statement is correct, i.e. that if n is not a multiple of 3, then n^2 is not a multiple of 3. So, as we are assuming that n is not a multiple of 3, we know that $n = 3k + 1$ or $n = 3k + 2$ for some $k \in \mathbb{Z}$. We will consider each case in turn. Suppose $n = 3k + 1$, then $n^2 = (3k+1)^2 = 9k^2+6k+1 = 3(3k^2+2k)+1$, which is not a multiple of 3. Now we consider the second case: $n^2 = (3k+2)^2 = 9k^2+12k+4 = 3(3k^2+4k+1)+1$, which is not a multiple of 3. So, in both cases n^2 is not a multiple of 3, and we have established the contrapositive, which is equivalent to the original statement.

B Arguments used in Experiment 2.

B.1 *The direct argument.*

Claim. The product of diagonal matrices is diagonal.

Argument. Let $A = (a_{ij})_{n\times n}$ and $B = (b_{ij})_{n\times n}$ be two diagonal matrices of size $n \times n$. Let $C = AB$. Then

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{in}b_{nj}$$

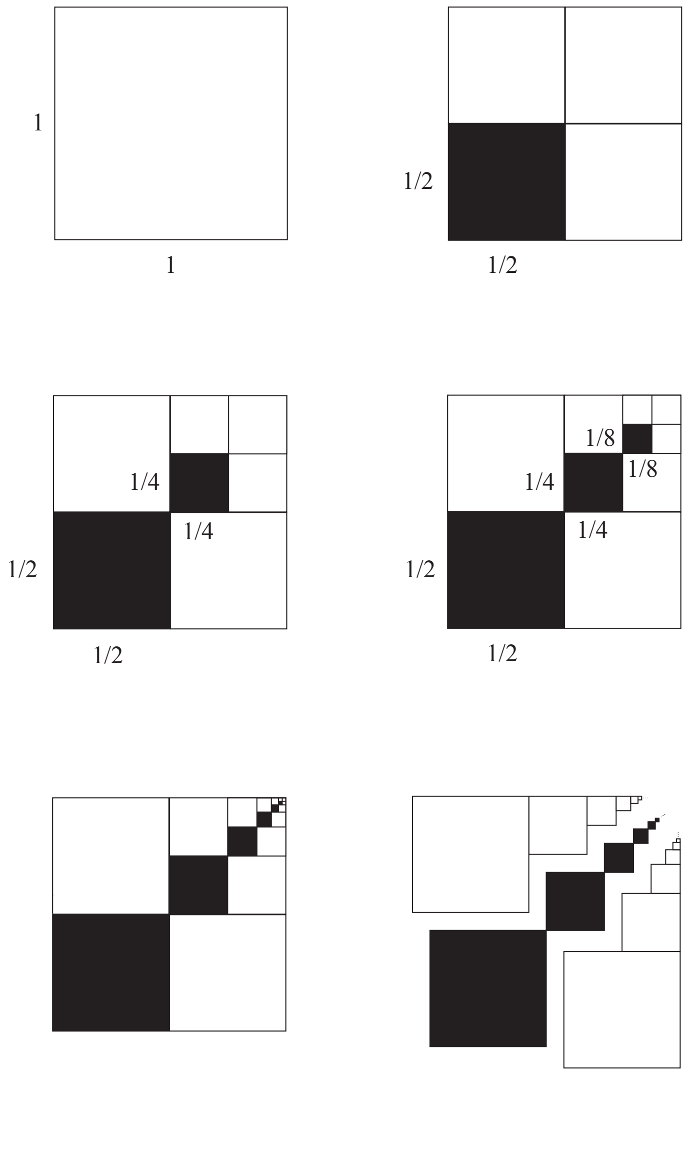

Fig. A.1. The visual argument for the claim $\sum_{n=1}^{\infty} \frac{1}{2^{2n}} = \frac{1}{3}$.

Since A is diagonal then $a_{ik} = 0$ whenever $i \neq k$. Therefore $c_{ij} = a_{ii}b_{ij}$. Since B is diagonal, $b_{ij} = 0$ whenever $i \neq j$. Therefore, $c_{ij} = 0$ whenever $i \neq j$. Thus C is diagonal.

B.2 The visual Young's argument.

Claim. Let ϕ and ψ be two continuous, strictly increasing functions. Suppose $\phi = \psi^{-1}$ and $\phi(0) = \psi(0) = 0$. Then, for $a, b \geq 0$, we have:

$$ab \leq \int_0^a \phi(x)\,dx + \int_0^b \psi(y)\,dy$$

with equality if and only if $b = \phi(a)$.

Argument. [Figure B.1].

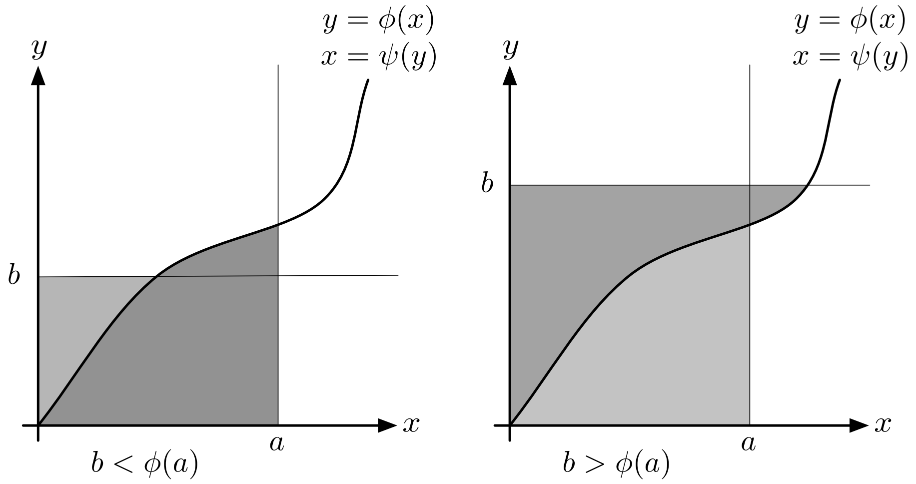

Fig. B.1. The visual argument for Young's inequality.