Simple categorization of mathematical objects: examining students' decisions

A report on the thesis of David Janda, May 2020

Adrian Simpson
Professor of Mathematics Education
Durham University

This is an interesting thesis describing the thoughtful development of a main study in to how people use a new concept acquired through one of two routes to categorise new stimuli as instances or non-instances of that concept. It fits well in the more basic cognitive tradition of mathematics education and thus attempts to adopt the methods of that tradition, particularly in utilising a more experimental approach to the work. Of course, the thesis is presented in what is not the candidate's first language which is immediately impressive, all the more so that this very rarely shows through: from a language perspective it is remarkably well written with surprisingly few grammatical or typographical errors.

Structurally it follows a fairly standard format, setting the work in the context of some key theories and existing relevant research, discussing pilot studies, the main study and drawing conclusions which bring us back to the theory and existing literature nicely. Where there is a structural issue is in the lack of a clear methods chapter. While many elements of method were covered in the pilot study chapter, where methods changed between the pilot and main studies, the new methods were not examined in the same depth. This is somewhat problematic because the strength of the main study lies in the data from the new methods - the qualitative data was much more revealing than the quantitative.

The theoretical considerations were generally dealt with well, albeit that I would like to have seen a greater depth of critical analysis. The thesis is grounded in three approaches to concept: process-object theories, image/definition and the historical perspective. In the end, only one of these is genuinely useful here - the image/definition distinction accounts well for the data observed. While reference is made at the end to the inappropriateness of process-object theories as mechanisms for accounting for the data (about which I would like to have read more), the thesis does not close the loop on the historical perspective. I would be interested to know what the candidate thinks about the adequacy or otherwise of Kvasz's re-coding, relativisation and reformulation ideas for the data in this thesis.

One of the strengths of the thesis is the core task: the categorisation of the Tall function. The Tall function itself is a clever idea, fitting with notions such as Dahlberg &

Housman's fine function (which I was surprised not to see included in the discussion), in that it delineates a clear subset of familiar objects, but is accessible both via definition and imagery. The classification of stimuli into examples and non-examples is a sensible approach and through the pilot study chapter we can see the candidate clearly developing their methods, analysing their approaches and polishing procedures. Some methodological aspects were lacking: as noted above, there was a lack of discussion about the qualitative methods, but there was also only partial discussion about quantitative analysis. While careful thought was given to crucial elements like how best to obtain a point summary of a set of reaction times, some of the bigger issues such as how to determine if two sets of reaction times were sufficiently different to draw conclusions about cause were not addressed. This led me to wonder whether a number of issues around sample size, lack of statistical testing, omission of variance measures etc. might have been avoided if more time had been spent on these broad methods issues.

The transition to the main study not only brought with it some new methods around the core task - one of which proves to be the major strength of the thesis - it brought a confusing extra which could have been omitted from the thesis without negative effect. It was never clear why the injective function classification task was added. I could see a number of plausible reasons for it - at the most basic, it might have given a baseline response time against which to reference participants' response times for the main task. That is, one could look at their Tall function response time *relative* to their average injective function response time which might then increase the power of any statistical analysis (which might help compensate for the low power given the sample size). There are other plausible reasons for looking at a second task, but none is clearly given here and the analysis and conclusions related to this extra task do not contribute to the argument.

In general, the quantitative analysis did not contribute as much as it could and certainly did not provide the insight of the qualitative analysis. I was surprised to see little statistical testing, and to see many descriptive summaries instead. Despite the low power of the sample size, I would expect to see differences like the between group response times on the first trial to be highly significant (they were highly dissociated according to graph 11). I would also have expected to see data summaries not just to include averages, but measures of spread as well. Having noted this, the clear difference between the two groups in terms of response rate and the impact of the second trial (in slowing down group I and speeding up group D) is fascinating and worth exploring in more depth. I am not convinced by the account given for this in chapter 4 and would like to have seen the read in terms of access to dual processes: access to a fast process speeding up group D and access to a slow process slowing down (and making more accurate) group I.

Some of the other quantitative analysis was less well focussed. For example, it was never very clear what the analysis in terms of number of changed decisions (3.5.2) was helping to uncover.

The key strength of the thesis, for me, came in the qualitative analysis. Despite the lack of discussion of method here, the analysis of the participants as cases contained real insight: we saw instances of the development of personal concept definitions and the delineation of sets of features which appeared to be in use to make categorisation decisions. It would be fascinating to think about these in terms of prototypes vs mini-theories. While the thesis notes that the participants do not seem to be using prototypes (perhaps because they have insufficient opportunity to develop one which might be sufficient to the classification task) I would argue that some could be seen as mini-theories. In particular, the ‘misty CI’ idea might be seen as people developing a complex classification tree or collection of disconnected potential diagnostic features of Tall functions, in contrast to those who seek an overarching principle. This is potentially very interesting indeed, but not addressed in much depth in chapter 3 and not at all in chapter 4. On balance, the thesis might have been stronger if the qualitative had been the focus, with a tighter discussion of quantitative in a supporting role.

I have attached below comments on each chapter highlighting where I see the key strengths and weaknesses. In summary, I think there is much to commend about the originality of the thesis: the task itself is clever and the approach taken is basically sound. The qualitative data particularly has the potential to ground a discussion about different ways in which learners might approach categorisation tasks that might be a valuable and publishable contribution to the field.

Appendix: Comments on chapters

Chapter 1

This was an interesting chapter covering a very wide range of relevant topics, though the extent to which it engaged deeply with those topics varied considerably, from the rather superficial to some delightfully deep discussions. The writing throughout was engaging, but I am concerned at the lack of a clear narrative structure acting as a thread through the piece. The use of a clear introduction and summary section helped with this, but there were many disconcerting moments when, as a reader, I could not see where a piece of argument linked to the pieces around it.

At its best, the depth of engagement with the topics was superb. The candidate's work comes alive when he is focussed and goes in to depth rather than breadth. I was fascinated by some of the thought behind the thesis, even at this early stage. The distinction between a binary operation as a process and an object (p 19) is really interesting: clearly a pupil will meet addition as a process repeatedly throughout their educational career, with only a very small handful of learners ever coming to see addition-as-an-operator as an object. Similarly, the discussion of categories is outstanding, well structured and insightful (and rather at odds with his discussion of objects and concepts which skates over the surface of too many issues). His linking together Wittgenstein, fuzzy sets and prototype theory works particularly well and shows someone capable of the sort of analysis and synthesis expected at this level. There is an equally deep, well written and insightful section on the development of mathematical concepts. This both shows the candidate has a good grasp of the history of mathematical ideas (for this level) and can link ideas from philosophy, education and mathematics together. That said, I would like to have seen this section more clearly linked to those around it - I think the argument is that we can learn something about personal concept development from society's development of the concept, but there is no connecting passages that link this section to those around it.

This last point is a key negative in this chapter - that the writing (at all levels) was not well structured. At the lowest levels of granularity there are occasional stubs of ideas which would benefit from expansion - e.g. "Similarly, the term *personal concept image* is used in the literature" (p13). I would have been interested to know what this meant and what literature this had come from and, ideally, seen the idea synthesised with the other ideas and analysed against them (for example, to what extent can one have a concept image which is not personal?). Or, as another example, "However, some of their conclusions have been questioned by some other researchers (Iannone et al., 2011)" - how have the conclusions been questioned, what do Iannone et al. have to say and where does the candidate stand on the apparent difference of views expressed? Some of these 'stubs' (where a train of thought is begun, but not developed) might be better removed - for example (though it pains me to say it), the idea of natural, conflicting and alien (p15) is not really relevant here (or at least the candidate does not demonstrate its relevance).

We also saw this at medium levels of granularity: a paragraph on Fischbein's figural concepts was in the section on examples, rather than concept image where it might have had a more obvious connection. Similarly, ideas about concept image/definition are scattered throughout the chapter when with more careful construction (and perhaps more time!) they might have been drawn together. And we saw it at the largest levels of granularity: I was surprised to see a discussion of methods for investigating concepts and categories in this chapter. While interesting, it would be better tied to the methods chapter: issues about how to deal with outliers in reaction time studies don't really belong in a chapter on mathematical concepts. Moreover, it discussed the methods for investigating categories before the thesis had dealt with what categories are.

The chapter covers a considerable amount of important literature and some key ideas which are crucial for the development of the candidate's thesis: concepts, objects, definitions, images, examples etc. The candidate highlights key points and draws out some fascinating issues - for example the idea of the use of examples as seeds of generalisation. There is a particularly nice discussion about the counterexample's status as relative to a statement or assertion (the candidate uses 'theorem' though of course a theorem cannot have a counterexample) and notes that this makes a counterexample quite different in status from, say, a generic example. Similarly, the candidate draws a link (justifiably hedged) with concept image/definition, Poincaré's view of Riemann/Weierstrauss and System 1/2 thinking which is novel and which I would have liked to have seen extended. It is important not to be afraid to challenge and I would have like to have seen more evidence of critical engagement with the literature, for example Watson and Mason's definition of 'example' is clearly capable of inappropriate extension. If "anything used as a raw material for intuiting relationships and inductive reasoning" is an example, then Pythagoras's theorem is an example of a triangle! Clearly an example has to be an instance of the concept or principle one is attempting to exemplify.

There are a few areas where I would want to challenge what the candidate is saying and seek exploration of his understanding. For example, it is not clear to me that a mathematical object is a form of declarative knowledge. Sternberg's definition ties to its linguistic origins as a fact "which can be stated" (i.e. can be declared) and it is not clear that an object (either a platonic object, a socially agreed object, a personal psychological object or even a physical object) can be declared. Indeed, one might argue that objects are not, of themselves, knowledge - one can have knowledge about an object (something one can state about it or something one can do with/to it - which correspond to declarative/procedural knowledge) but the object itself is not knowledge. It may have been useful to draw this out by exploring how theories like Dubinsky's form objects from procedures - which is alluded to on p12 - where their mental objects are 'schema': something like a container of facts and processes which are linked together in some way.

The candidate twice draws Dubinsky, Sfard and Hejny together as if they are closely related. While the case is made for Dubinsky and Sfard being process-object theories (often also connected to Grey and Tall's procept theory), Hejny's seems rather different

in that it emphasises the models, rather than the processes, as the genesis of the object. Again, drawing those distinctions out would have been fascinating.

The chapter ends with a clear summary which culminates in a statement of the research question. The genesis of the research question is thus very clear, though the question itself - as written - is not particularly well formed: it admits only a yes/no answer. The thesis appears to be more concerned with the ways in which categorisation task responses might provide insight into CI/CD rather than whether such responses provide insight.

Chapter 2

To some extent, this chapter is incorrectly titled: it really concerns the pilot studies, rather than the research design (which changed quite dramatically for the main study). It would have been better to make that clear with the title and then have a separate chapter covering methods for the main study (possibly referring back to overlapping methods from the pilot study where appropriate). The progress through the chapter made narrative sense, once the reader understands that it is effectively about piloting. The opening of the chapter gives a very clear breakdown of the research question into methodological considerations - the need to identify the mechanism of categorisation and the need to identify the form of participants' CI/CD.

However, the chapter contains a crucial methodological approach which is used in the main study: that reaction time gives insight into the categorisation strategy. However, this is simply stated - there is no justification given for this and the only links to the contents of the cited literature which might explain such a link lie buried in the misplaced section in chapter 1. There is a need to examine this in much more depth: the thesis is relying on two key pieces of research, Ashby, Boynton & Lee (1994) and VanRullen & Thorpe (2001). The former does indeed suggest a link between reaction time and mental structure (the longer reaction time indicating something being closer to the boundary of a category) but, if anything, VanRullen & Thorpe's study suggests problems with using reaction time to distinguish forms of categorisation - their study undermines the assertion that natural categories are classified more quickly than artificial ones. Given the importance of reaction time to the study, this issue needed much more exploration.

Given the pilot nature of the two studies discussed, they appear to have been appropriately carried out and some of the analysis is interesting, though there are some key issues which were worthy of further discussion. For example, the near total dissociation of examples and non-examples in terms of response time (as seen in graph 6) is stark and dealt with very abruptly (with the assumption that a non example can be confirmed with a single feature, while examples need all features checked).

The first pilot study was clearly helpful in addressing technical issues and the thesis outlines many of these. With more space, I would like to have seen more discussion about potential causes and solutions to these issues. For example, the first figure(s) effect does seem to have an obvious plausible account: that it is a form of the practice effect which can be addressed by starting the activity off with a few simple (and obvious) practice examples/nonexamples which help participants get used to the experimental set up, the task and the concept and can also act as a filter (for example, one could exclude anyone who scores below a fixed level on the simplest of examples).

Pilot study 2 was not well designed to address the issue of the awareness of the reaction time testing - it would have been much more sensible to have a (random) half of the participants undertake tasks 2 and 3 in one order and the other half in the other order. As it is, the fact that the participants classified the items more quickly when they were aware of being timed could equally well be attributed to classifying items they had already seen.

There were a number of issues about scholarly presentation standards. For example, it would be usual to report statistical results in an accepted format, e.g. on p.39 to include the degrees of freedom in the expression (thus $t(21)=10.096$, $p<0.0001$); in graph 1, I would also expect axes to be labeled with units and the caption to contain more information; in graphs 5 and 7, I would have expected to see the line of best fit (or two if it was worth analysing the two groups of questions separately) and in the text to have seen a statistical test for the significance of the relationship; in most of the graphs I would expect to see indications of variance.

There was a careful discussion about how to aggregate and summarise the data from a statistical viewpoint, but no discussion about the conceptual issues: when we aggregate data across participants to draw conclusions about concept images we make assumptions about a consistency of images across people, while when we aggregate across stimuli we make assumptions about consistency of stimuli. The latter are open to examination (we know already what the stimuli are) while the former are not. This suggests being wary of aggregating across people without care.

Chapter 3

This chapter is clearly the meat of the thesis. There was much that was interesting and at times really thought provoking, though the balance of the chapter might have been altered to allow a stronger focus on the qualitative over the quantitative, as it tended to be the former where the most interesting material lay. This chapter is probably the place where the lack of a clear methods section is most sorely felt: while some methods followed naturally from the pilot studies, some did not and were not explored or fully justified. The chapter might also have benefited from a summary - or even perhaps one

at the end of the two main parts - to allow the reader to keep track of where the argument is developing and which research questions have been explored.

There were a large number of research questions introduced. Some of these followed naturally from the pilot studies, but there was little discussion of the need to introduce a new set of stimuli (injective functions), the new activity (images) and the new form of data gathering (interviews). As is often the case with doctoral research, candidates can get in to the mindset that more is better when it comes to data, whereas it is often better to focus more carefully on less data, collected according to a clearly defined plan. The image activity was new and very different from the pilot study activity. It was also rather unusual because some of the choice of images used to describe the Tall function set was in the control of the participants, rather than the researcher. It would have been useful to see a discussion about the reasoning behind this method. Similarly, the rationale for the injective function categorisation task and the interviews needed to be explained. As it was, the injective function task and analysis played little role.

The chapter splits neatly into two sections: quantitative and qualitative. The latter is by far the more interesting and it is to these sections where I think potential journal papers might come more easily. The quantitative analysis suffered from some of the problems of the pilot studies (lack of presentational quality in the graphs, including no indication of variance etc.) and the whole of this section would have benefited from a clearer tie to the research questions. Perhaps structuring the chapter to address the research questions one by one would have helped, or tying each result to the question it was intended to answer. For example, it is not clear what question is addressed through the analysis of the number of changes made between the two categorisation tasks. In addition, structurally it may have made more sense to portray the first three students as a further pilot study, and incorporate it into the pilot study chapter. This might have allowed more discussion of the rationale for the new elements in the main study and addressed some of the issues here. It might also have allowed for a discussion of the analytic methods, particularly in respect of the interviews.

In terms of quantitative methods, there was a clear justification for the choice of sample. I would have liked to have seen some discussion about sample size - if not a power analysis, at least some indication of why 18 participants sufficed for the research design. For a pure randomised controlled trial perspective, the sample would have an 80% chance of detecting a real effect of over half a standard deviation difference in normally distributed data - that would have to be a very clear effect. It is also not clear why there were 7 participants in group D and 11 in group I - random allocation could have been done to ensure equal group sizes. I was surprised to see little statistical analysis of the results - given the decision to randomise to groups, the opportunity existed to reasonably exclude group selection effects as a sole cause of observed differences by checking if between group differences were significant or not. Again, the lack of indication of between participant variance made it impossible for the reader to do this. In some of the analysis, it may be worth thinking about whether these was a ceiling effect, particularly in discussions of the between trial differences: a number of people

scored very close to 100% and so the room for improvement was much reduced compared to those who scored closer to chance levels.

There was a set of very interesting findings in relation to reaction times: that those with images not only responded much more quickly than those with definitions, but adding the definition slowed the former group down (albeit resulting in more accurate categorisation) and adding the images (or a simple familiarity effect) speeded the second group up (with slightly improved performance). This is clear at the individual level in graph 11 - every participant in group D was the same speed or quicker and every participant in group I was the same speed or slower. It would have been good to draw this out as explicitly as this. However, I was not convinced by the examination of MIN-MAX here; while the MAX might reasonably represent the response for the most difficult to classify stimulus (for each individual) and is thus worth exploring, the difference represents only the range of response times and would be completely overwhelmed by the MAX value. It is interesting to speculate why the I group responded more quickly on the hardest to classify both in the first and second trial (though that might be accompanied by whether they classified correctly!)

The qualitative section was much stronger, though it was unfortunate that it did not examine the qualitative analytic methods in any depth. There was a clear introduction to the section, outlining what was to follow and the main part of this section was a tour de force. There is a very interesting discussion throughout about the features which the participants were attending to. For example, about the nature of +/- infinity and whether the respondents allow them within their decision making process. I also found the *range* of features to which the respondents attended to be fascinating. Pages 62-72 are very powerful - albeit I think the candidate could have done more to classify the different responses more clearly and could go in to more depth. There were further intriguing ideas which were not fully followed up, for example, the determining vs non determining aspects, the disclaiming aspects etc. were discussed in a brief, interesting section, but not well illustrated and not fleshed out.

There were many interesting elements worthy of discussion. It is interesting to note, following R266's definition, that all of the examples of a Tall function (in both test and training phases) have positive maxima and all of the non-examples have positive or zero (or positive infinity) suprema to the function's range. So the respondent's definition is entirely consistent with the image classifications they have been given. Equally, I appreciate seeing the two respondents who were able to give concise personal concept definitions, but I was left wondering how many people were able to give such definitions, how accurate they were and whether the candidate was able to discern any key features about the kinds of personal definitions which were constructed. Similarly, I found the issue of the 'misty' concept image fascinating - the distinction between a participant trying to form a definition from a large set of decision rules and a participant looking for a simple unifying principle is really interesting. I would have likes to have seen that types of distinction drawn out and some of the words of R235 transcribed in the same way as the more unified definition responses. It was useful to look at individual respondents in section 3.7, though again I wanted to see more depth. I did not

see the value to 3.7.4 - nothing of interest seemed to come from this analysis. It would have been interesting instead to look at the difference between the first and second interviews - how did someone's approach to categorising change when they were given the concept definition from when they had only had a concept image to work with?

In terms of readability, it would have been useful to code the transcripts so that a reader could know at a glance which group the respondent was in and from which interview the extract was taken. For example, I had to flick a long way back to find that respondent 234 (p65) was in group D, but I did not know whether their discussion about the nature of the maxima or minima came before or after the image teaching session.

On balance, I think the chapter would have been more valuable overall if the quantitative elements had been much reduced and more space given to a deeper and more systematic study of the interview data.

Chapter 4

This chapter usefully brought the reader back to the research questions and linked the findings to the literature. I particularly liked the explicit referencing back to the research questions. At its best, i could see the way the conclusions were justified by the data in the previous chapter, but this was not always the case. It is interesting to note that some of the most substantial results are tied back to the evidence of the qualitative analysis, which links to my comments on chapter 3.

I am not sure i am convinced that the data supports the conclusion in 4.1.1. It is important to distinguish between the cognitive strategies open to the two groups. Group D were given a definition and while they could have used that to form a visual, holistic concept image, it is reasonable to assume that they would disproportionately access systematic feature checking processes as decision makers. Group I in contrast were given images and while it is possible (though arguably unlikely) that they formed a personal definition, it is reasonable to assume that they would disproportionately access holistic similarity processes as decision makers. The timing and accuracy would appear to be consistent with such a theory. However, differences in timings between people has not been examined in that much detail in chapter 3, and they would it be able to distinguish between those who are simply faster and slower for the same cognitive process - two people who use their concept image to make a decision could still vary considerably in the length of time taken to make such a decision. Unless one could convincingly show a very substantial dissociation between image and definition based processes, one would find it difficult to make conclusions about cognitive preferences (from the quantitative data of this sort). Similarly, I am not convinced by part of the conclusion of 4.1.6: while it is clear that the two groups come to around the same average reaction time after their second trial, it would not seem to be reasonable to conclude that this is a result of the development of the concept image, but by the development of a dual approach: in trial 1, group I might disproportionately access the

(fast) holistic process and group D the (slow) feature matching process; in trial 2 both groups might have access to both processes and apply them in the way system1/2 theory would suggest - roughly speaking, use the fast process for the simple cases and the slow process for the more difficult ones. That is consistent with group I getting slower and group D getting quicker and both groups converging on similar speeds.

As mentioned in chapter 3, there was little discussion on the injective function task and the analysis was rather superficial, so the conclusions in 4.1.7 were difficult to see as justified.

The links back to the literature were often clear and insightful, particularly around the concept image/definition literature. The chapter makes an important point about the way process-concept theories do not allow us to account for the concept formation in this case - the stimuli in this case would all be expected to be objects already for these participants and the concept is, in effect, a subconcept of something already familiar. This leads to two questions: why did the literature review cover process object theories and whether the distinction discussed above between Dubinsky/Sfard/Grey&Tall theories and Hejny's could be important here? The chapter makes a point of stating that key strength is the way the work drew on CI/CD, process-concept and historical development ideas; but the last of these from Kvasz was missing from the chapter, which is a pity.

I was pleased to see explicit reference to the study's weaknesses and limitations and an attempt to look at the implications for practice.