SYNTACTIC AND SEMANTIC ITEMS IN ALGEBRA TESTS – A CONCEPTUAL AND EMPIRICAL VIEW

Reinhard Oldenburg, Jeremy Hodgen, Dietmar Küchemann

Goethe University Frankfurt, King's College London

In this paper, we draw a distinction between syntactic and semantic aspects of algebraic thinking. We examine the hypothesis that these two aspects can be distinguished empirically using test items. We present exploratory analysis of a test of algebra based on a large sample of students aged 11-14 in England and contrast this to a previous analysis of older German students (Oldenburg, 2009). This analysis indicates that there are considerable difficulties in operationalizing the distinction using test items, but suggests a potentially fruitful line of analysis may be to treat the semantic aspect as consisting of two sub-dimensions, based on whether one or many meanings or interpretations appear to be required.

INTRODUCTION

Although research on algebra education is now at a mature stage, many aspects of student progression and learning remain poorly understood. Of particular interest, is the relationship between the syntactic understanding of the rules and procedures involved in the manipulation of symbols and the semantic understanding required to interpret and attach meaning to symbols and rules. In a previous study, Oldenburg (2009) argued that test items could be constructed to measure and distinguish these two aspects of algebraic thinking. Using this distinction, he found that students can gain some level of proficiency in one aspect while being weak in the other, but that both aspects were necessary for a sophisticated understanding to develop. Whilst accepting the utility and validity of the syntactic / semantic distinction both to describe algebraic thinking and to inform pedagogy, the other two authors of this paper were sceptical about whether this distinction could be applied to items and whether such thinking could be measured using test instruments. The present paper is a result of the authors' subsequent debates and explores the meaning, usability and limitation of this pair of constructs in depth. To do this, we analyse the performance of different items in an algebra test administered to a nationally representative sample of students in England.

THEORY: DISTINGUISHING SYNTACTIC AND SEMANTIC TASKS

If we consider algebra as a language, then, as with any other language, algebra can be thought of as having syntax and semantics, which must be applied to any use of the algebraic language, e.g. in communication and problem solving. [1] Since thinking is to a large extent based on language, syntax and semantics can be viewed as aspects of thinking and understanding in general. In general, any algebraic thinking involves both aspects and hence distinguishing the two is not straightforward. Nevertheless, we contend that some algebraic tasks are more amenable to syntactic approaches,

whilst others are more amenable to semantic approaches. For the purposes of the present paper, we give the following working definition:

A syntactic task, or assessment item is one that can be solved by actions triggered by the syntactic structure of the expression alone without involving a mental model for interpretation, i.e. without having mental objects referred to by the symbols. For example, it is possible to solve the expansion of $(x+y)^2$ by purely syntactic thinking, because the pure lexical structure may activate the schema of the binomial theorem.

A semantic task, or assessment item is one in which the need for the interpretation of symbols (i.e. the construction of a mental model of objects denoted by symbols) is dominant in successful solutions. For example, in order to give a general expression that allows one to calculate the number of wheels a certain number of cars have, one has to activate semantic thinking to symbolize the number of cars by a letter and relate this letter in its domain of interpretation with the wheel number.

Syntactic tasks by this definition are those that can be successfully carried out by "term rewriting systems" as defined in computer science (Baader and Nipkow, 1999). However, such systems cannot carry out tasks that require students to link the algebraic language with objects and concepts from outside mathematics. We contend that tasks that require at least a complete sentence in natural language are very likely to be semantic. We note that this definition is based on the anticipated processes that will be used when solving the items. Thus, in order to classify tasks as mainly syntactic or mainly semantic, one needs hypotheses about the way the tasks are likely to be tackled by students.

We do not claim that our distinction between syntax and semantics is entirely unique, although other researchers' definitions differ in some key respects. Most recently, the distinction has been used to analyse different approaches to, and aspects of, the production of proof (Weber and Alcock, 2004; see also, Iannone and Nardi, 2007). Weber and Alcock distinguish between "syntactic proof production", by which they mean "one which is written solely by manipulating correctly stated definitions and other relevant facts in a logically permissible way", and "semantic proof production", by which they mean "a proof of a statement in which the prover uses instantiation(s) of the mathematical object(s) to which the statement applies to suggest and guide the formal inferences that he or she draws" (p. 210). One problem with this definition is that the precise meaning of both "stated definition" and "permissible way" is rather vague. However, Weber and Alcock's approach shares with ours a concern for the importance of the domain of reference.

Kieran (1996) draws a distinction between 'transformational', or rule-based, activity, 'generational' activity in which the objects of algebra, expressions and equations, are formed, and 'global, meta-level' activity, which includes problem-solving and modelling. Whilst she argues that transformational activities are more often based on syntactic rules, she notes that they can in certain cases be legitimated by semantic arguments (e.g., transforming $1/(1-v/c)$ to $1-v/c$ in special relativity incorporates some

physical assumptions). Further, although Kieran argues that much of the "meaning-building" for algebraic objects takes place through generational activity, some generational activity can be largely rule-based (e.g., the generation of the polynomial sequence, $(x\text{-}1)^n$. Neubrand and Neubrand (2004) draw a distinction between technical, or procedural, tasks and modeling tasks, which often require semantic, or meaning-based, constructions. However, they note that technical calculation tasks may require some meaning-based activity, particularly those involving longer algorithms. Similarly, modelling requires students to operate on symbols as well as constructing meaning for the symbols. Although neither Kieran's nor Neubrand and Neubrand's dichotomies are completely identical to the syntactic/semantic distinction, their analyses do serve to remind us that the distinction between syntactic and semantic tasks is somewhat blurred, specifically:

- Some items can be tackled both syntactically and semantically. For example, expanding $5\cdot(x+2)$ may be done either by syntactic matching to the pattern of the distributive law $a\cdot(b+c)$ or by applying the semantic way of interpreting the expression as the area of a rectangle.
- A semantic approach may be helpful to help a learner self-correct the misapplication of syntactic rules. In the case of the expansion of $(x+y)^2$ referred to above, semantic thinking (by, for example, substituting some numbers for the symbols) can help avoid the common expansion error of x^2+y^2.
- Even items that are mainly semantic in nature (e.g. "Explain the meaning of $2g+4r$ in some context") involve at least the syntactic ability to read the expression. Here, one needs the semantic understanding that $2g+4r$ is a legitimate (set of) numbers/answer not just that the expression can be read as "2 times g added to 4 times r".
- Modelling tasks (i.e. generating an expression or equation to describe a situation or relation) require at least the minimal syntactic competence to write down the expression. So, while a mental distinction may be possible, it can be blurred as soon as the communication processes are required.

One could argue that the entanglement of syntactical and semantical aspects within the domain of algebraic thinking is clear from the outset, e.g. one could read Lins' and Kaput's (2004) definition of algebraic thinking that way: "First, [algebraic thinking] involves acts of deliberate generalisation and expression of generality. Second, it involves, usually a separate endeavour, reasoning based on the *forms* of syntactically-structured generalisations, including syntactically and semantically guided actions" (Lins & Kaput, 2004, p. 48). In fact, certain algebraic actions may be justified from semantical or from a syntactical perspective. Indeed we even more assume that this effect may depend on the development of algebraic thinking, i.e., we suggest that, as students' algebraic thinking becomes more sophisticated, the thinking evoked may change and tasks that previously required semantic approaches may be solved using purely syntactic approaches. Moreover, the hypothesis that items (and

not a particular student's way of doing an item) can (at least to some extent) be classified as either syntactic or semantic assumes that typical students have a preferred way of tackling these problems. We note also that it is possible a students' preferred approach may be strongly influenced by the teaching approaches adopted by their teachers. Hence, whilst it is possible theoretically to distinguish between syntactic and semantic aspects of algebraic thinking, it is not clear whether these aspects can be reliably distinguished empirically through test items.

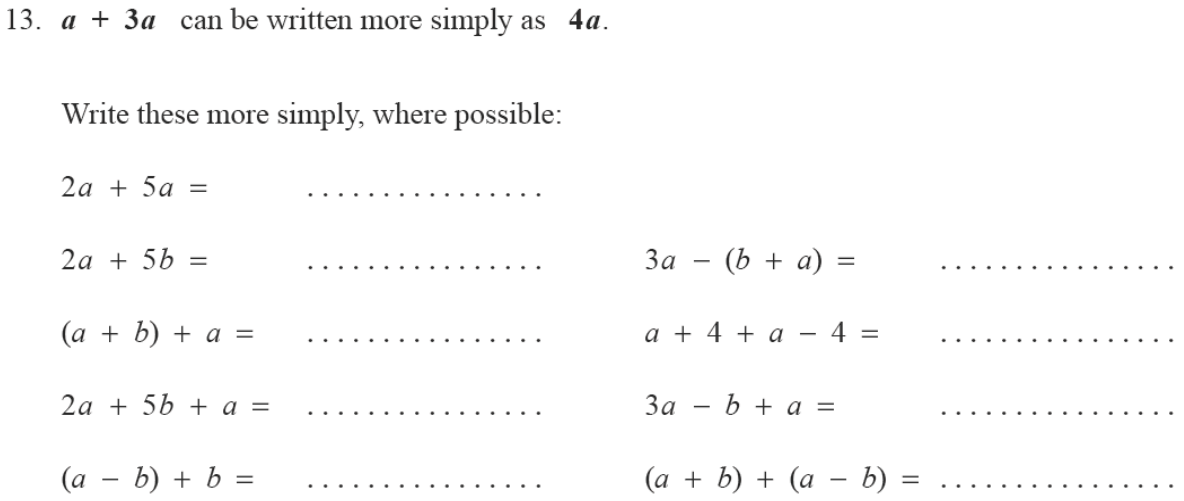
METHODS

In this paper, we analyse the performance of different items on an algebra test originally developed in the 1970s as part of the Concepts in Secondary Mathematics and Science (CSMS) study (Hart et al., 1981). In 2008 and 2009, these tests were administered to a nationally representative sample of 5115 students in England aged 11-14 as part of the Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS) study (Hodgen et al., 2009). These data were collected as part of a larger study designed to examine the difficulties that students in England encountered in algebra and multiplicative reasoning and was not specifically designed for this study.

The focus of the CSMS/ICCAMS algebra test is on generalized arithmetic (Küchemann, 1981). Drawing on, and extending, Collis's (1975) analysis of the different ways in which pronumerals can be interpreted, items were devised to bring out the following six categories (Küchemann, 1981): Letter evaluated, Letter not used, Letter as object, Letter as specific unknown, Letter as generalised number, and Letter as variable. We note that the test items were not developed specifically to address the syntactic / semantic distinction. Indeed, items in a category may involve syntactic or semantic approaches, although items in the two categories, 'letter as generalised number' and 'letter as variable', are more likely to be semantic.

The items were independently coded as syntactic, semantic or mixed by the three authors of this paper using only the definition of the categories given in the Theory section above, with mixed used for items felt to involve both aspect. Examples of classifications are given in figure 1 and 2. We assume that the classification in figure 1 can be easily agreed on, figure 2 presents items that are potentially difficult to classify. One may say that all of them have a syntactic appeal if attacked by replacing one sub-expression by a part that is known to be equal. This is what puts 5a,5b into the mixed category (as there are references to concrete numbers as well which constitute the semantical aspect). In the case of 5c, however, we assume that plugging 8 for $e+f$ in the second equation is different from tasks the student may have met before, so that according to the above definition ("A syntactic task, or assessment item is one that can be solved by actions triggered by the syntactic structure of the expression alone. ") we miss the triggering effect to take place ion the student that we assume to do the item.

The results of this initial coding exercise reflect the problematic nature of applying the syntactic/semantic distinction to items. The overall inter-rater reliability (3 rater, 3 categories) measured by Cohen's kappa is =0.54 which can, according to Landis and Koch (1977), be judged as moderate agreement. Inspection shows that this is mainly due to one rater who used the mixed classification a lot. Between the other two raters we find κ=0.70 which is a good inter-rater reliability. A final agreed classification was developed through discussion. In this final classification, of the 51 items in total, 18 were coded as syntactic, 25 as semantic, and only 8 as mixed.


13. $a + 3a$ can be written more simply as $4a$.

Write these more simply, where possible:

$2a + 5a =$ .....

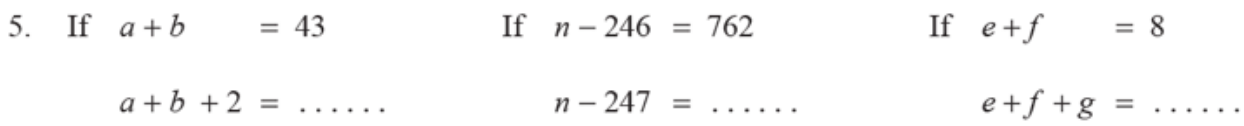
$2a + 5b =$ ..... $3a - (b + a) =$ .....

$(a + b) + a =$ ..... $a + 4 + a - 4 =$ .....

$2a + 5b + a =$ ..... $3a - b + a =$ .....

$(a - b) + b =$ ..... $(a + b) + (a - b) =$ .....

Figure 1: All nine items in Question 13 were coded as syntactic


5. If $a+b = 43$ If $n-246 = 762$ If $e+f = 8$

$a+b+2 =$ ..... $n-247 =$ ..... $e+f+g =$ .....

Figure 2: The first two items in Question 5 were coded as mixed; the third item ($e+f+g=$) was coded as semantic.

REVISITING OLDENBURG'S ORIGINAL FINDINGS

Oldenburg's (2009) original paper reports findings from a different algebra test performed with 11th graders (aged 16) in Germany. This test included many items from the ICCAMS test together with items on more advanced symbolic manipulation (e.g. simplifying square roots) and on real world applications (e.g. translating between real world contexts and algebra). [2] The test data from 2008 indicated a rather low correlation of r=0.33 between syntactic and semantic items. In the meantime this test has been conducted with many more students and higher values of the correlation in different groups up to r=0.54 have been found. From comparing these various groups one can deduce the rule that the correlation is higher in higher achieving schools. This result cannot be explained by a ceiling effect as even in the weakest group facility (i.e. the fraction of correct answers) on syntactic and semantic

items has been 33% resp. 53% so that no ceiling or floor effects that would reduce the correlation have to be anticipated.

One potential explanation for these low correlations is that it may be that any arbitrarily chosen groups of items of similar facilities would tend to have a similarly low correlation. In order to test this hypothesis, we performed a bootstrapping process on the data from Oldenburg (2009). [3] This produced a large set of randomly chosen scales with a higher average correlation of 0.64 (standard deviation 0.062). We conclude that the selection of syntactic items is not arbitrary.

EXAMINING THE PERFORMANCE OF SYNTACTIC AND SEMANTIC ITEMS

We now consider our analysis of the 2008/9 ICCAMS dataset. The scales or item groups for both syntactic and semantic groups were found to have good internal reliability: Cronbach's alpha was 0.87 for the syntactic scale and 0.86 for the semantic scale. However, in sharp contrast to the original study, the correlation between the two scales was r=0.75 rising to r=0.78 (when blank responses are treated as incorrect rather than missing).

First, the difference may indicate some sample dependence. For example, the relationship between syntactic and semantic approaches may be influenced by the different curricula and pedagogic approaches in England and Germany. Second, as described above, we aimed to code as many items as possible. It is clear that this approach may tend to classify 'mixed' items as either syntactic or semantic. This may be reflected in the difficulties that we encountered in the coding process. Third, unlike Oldenburg's (2009) study, the ICCAMS test items were not specifically designed to assess the syntactic / semantic distinction. In fact, as we have already noted, the ICCAMS test was not specifically designed to examine the syntactic/semantic distinction and sought to use items that use "problems which were recognizably connected to the mathematics curriculum but which would require the child to use methods which were not obviously 'rules'." (Hart and Johnson, 1983, p.2). As a result, the test items may be biased towards syntactic items that require some sematic thinking.

To investigate the second and third of these points, we reduced the syntactic scale to a core of 9 syntactic items originally coded as syntactic by all three authors. This reduces the correlations to 0.66 (or 0.69 with missing values treated as incorrect). However, a similar process of elimination from the semantic scale does not bring the value further down, so that still the explanation is not satisfactorily. We propose an explanation that is related but not identical to aspects and uses of letters as symbols in algebra.

In mathematical logic (cf. Tourlakis 2003, p. 53) semantics is defined using interpretations. An interpretation of a set of formulae of predicate calculus is given by a set S (the domain of the interpretation) and an assignment that gives an element of

S for every occurrence of an unbound variable in the formula, and functions and predicates over S for every function and predicate symbol in the formulae. After applying all these assignments the formulae reduce to statements in the domain with no unbound variables remaining. For working with specific numbers it is thus enough to consider one interpretation, but to prove that a formula is a tautology one has to show that it is true in all interpretations. This distinction is crucial for learners. Küchemann (1981) has carefully drawn the distinction between an understanding of letters as specific unknowns and a more sophisticated understanding of letters as generalized numbers or variables. We suggest that the former, although semantic, requires the learner to consider only one possible interpretation or meaning, while the latter requires one to consider multiple interpretations. This suggests that treating the semantic scale as one-dimensional may not be justified. Thus we split up the semantic scale into two subscales:

- **SES – *Semantics with single interpretation***: These are items that require only one interpretation to be considered. Thus they can be solved by mentally replacing the x by one number. A key example is the item: Write down the smallest and the largest of these: $n + 1$, $n + 4$, $n - 3$, n, $n - 7$
- **SEM – *Semantics with multiple interpretations***: These items can only be solved if multiple interpretations of the variable are considered. A key item is: Which is larger, $2n$ or $n + 2$?

This gives the following correlations:

r	**SYN**	**SES**	**SEM**
SYN	1	0.87	0.50
SES		1	0.55
SEM			1

Table 1: Correlations between syntactic and two kind of semantic items

If one replaces SYN by the core of transformational items then the correlation SYN-SEM goes down to 0.45 and of SYN-SES to 0.71. Again with the standard definition of SYN, one observes an interesting dependency of the correlation SYN-SEM which is for students aged 12, 13 and 14, respectively, 0.45, 0.47, 0.54. This is consistent with the observation made above that the test from Oldenburg (2009) shows higher correlations in better performing schools. Our interpretation of this is that it is possible to gain a certain level of facility with either syntactical or semantical items without understanding the complementary aspect, but that higher achievement levels (as typically reached by older students) generally require a tighter integration of both aspects.

This suggests – although one should keep in mind that this claim has no broader base than the one presented –that a simple dichotomy between semantic and syntactic item may be useful as a first approximation but blurs some important distinctions. The three scales defined above seem to better model the cognitive structure of students.

In order to examine this hypothesis further, we performed a structural equation modeling of the situation with SYN, SES and SEM as latent variables. Fitting such a model enables the identification of individual items that do not fit well into a scale. [4] The resulting model showed a good model fit (RMSEA=0.047, CFI=0.37). See Figure 3.

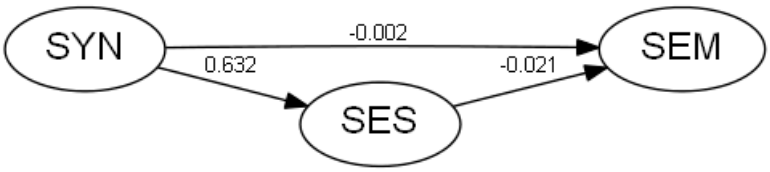


Figure 3: The structural model of three latent variables

The model equations were as follows:

$$SES = 0.63*SYN + error \quad (1)$$

$$SEM = -0.002*SYN - 0.011*SES + error \quad (2)$$

Equation (1) suggests that there may be a bridge from understanding the syntax of expressions to the ability to use it with specific (although unknown) numbers in one interpretation. One might conjecture that applying the interpretation (of letter as a specific but unknown number) requires at least the syntactic ability to plug in values for letters and evaluate the result. Equation (2), on the other hand, suggests that a fuller semantic understanding may develop relatively independently of syntactic and semantics with a single interpretation. [5]

CONCLUSION

As we have seen, there are considerable methodological issues in operationalizing the distinction between syntactic and semantic modes of thinking in existing test items. We found the process of coding the items to be problematic. This difficulty is reflected in the theoretical literature where the two modes are not treated as entirely distinct. Our initial approach was to attempt to code as many items as possible as either syntactic or semantic. On this basis, our analysis of the English data suggested that these data do not show the same pronounced distinction between syntactic and semantic items as was previously reported for the German data (Oldenburg, 2009). We have hypothesised possible reasons for this, such as some interdependence with different country curricula. In addition, the two samples were of different ages and it may be that the classification of items as syntactic or semantic may be age-related. Our analysis also suggests that the semantic scale may not be uni-dimensional. By classifying the semantic aspect into two sub-dimensions, based on whether one or many meanings or interpretations appear to be required, we found a relationship between syntactic and semantic items based on a single meaning, but not between either and sematic items requiring multiple interpretations. We emphasise, however, that this is exploratory analysis and further studies are needed to explore these questions deeper. Nevertheless, our analysis suggests that it may be possible to

identify syntactic and semantic abilities using test items. We note, however, that this may require items designed specifically for this purpose.

NOTES

1. Pragmatics is also considered to be an important aspect of language (Rowland, 199X). However, it is not considered here as it is beyond the focus of the algebra tests considered here.

2. In Oldenburg's (2009) original study, items were coded on a similar basis to the process described here by the sole author.

3. The bootstrapping was conducted as follows. We drew 10,000 samples of groups of seven and 30 items, exactly the same size as the syntactic and semantic groups in Oldenburg's (2009) study. Samples were retained if the average facility was within 15% of the average facility in the syntactic (33%) and semantic (53%) item groups and otherwise discarded. The average correlation was calculated on the basis of all the retained samples.

4. We used the R package lavaan using a polychoric covariance matrix (calculated with the R package polycor) in order to compensate for the binary indicator variable used here. The binary coded items were used as indicator variables with free weights that the three latent variables load on. The variance of the latent constructs was fixed to be 1.

5. The fact that both weights are negative is not important because both numbers are very small and don't differ from 0 significantly (standard errors are 0.026 and 0.021 for the first and second coefficient, respectively).

REFERENCES

Baader, F., & Nipkow, T. (1999). *Term Rewriting and All That*. Cambridge: University Press.

Engelbrecht, J. et al. (2005): Undergraduate students' performance and confidence in procedural and conceptual mathematics. *Int. J. Math. Educ. Sci. Technol. 36*(7), 701-712.

Hart, K., Brown, M. L., Küchemann, D. E., Kerslake, D., Ruddock, G., & McCartney, M. (Eds.). (1981). *Children's understanding of mathematics: 11-16*. London: John Murray.

Hart, K. M., & Johnson, D. C. (Eds.). (1983). *Secondary school children's understanding of mathematics: A report of the mathematics component of the concepts in secondary mathematics and science programme*. London: Centre for Science Education.

Hodgen, J., Küchemann, D., Brown, M., & Coe, R. (2009). School students' understandings of algebra 30 years on. In M. Tzekaki, M. Kaldrimidou & H. Sakonidis (Eds.), *Proceedings of the 33rd Conference of the International Group for the Psychology of Mathematics Education (PME 33)* (Vol. 3, pp. 177-184). Thessaloniki, Greece: PME.

Iannone, P., & Nardi, E. (2007). The interplay between syntactic and semantic knowledge in proof production: mathematicians' perspectives. *Proceedings of CERME5*.

Küchemann, D. (1979). Children's Understanding of Numerical Variables, *Mathematics in School, 7*, 23-26.

Küchemann, D. E. (1981). *The understanding of generalised arithmetic (algebra) by secondary school children*. Unpublished PhD thesis, Chelsea College, University of London.

Kyeong, J. M. et al. (2012). Semantic and syntactic reasoning in the learning of algebra. *ICME*.

Landis, J. R., & Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lins, R., & Kaput, J. J. (2004). The early development of algebraic reasoning: The current state of the field. In K. Stacey, H. Chick, & M. Kendal (Eds.), *The future of the teaching and learning of algebra*: The 12th ICMI Study (pp. 47-70). Dordrecht, The Netherlands: Kluwer.

Neubrand, J., & Neubrand, M. (2004). Innere Strukturen Mathematischer Leistung im PISA-2000-Test. In M. Neubrand et al. (Eds.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland.* (pp.87-108). Wiesbaden: VS Verlag.

Oldenburg, R. (2009). *Structure of algebraic competencies*. In V. Durand-Guerrier, S. Soury-Lavergne & F. Arzarello (Eds.), Proceedings of the Sixth Congress of the European Society for Research in Mathematics Education (CERME 6) (pp. 579-588). Lyon, France: Institut National De Recherche Pedagogique (INRP).

Oldenburg, R. (2010). A re-analysis of TIMSS data using Statistical implicative analysis. *Quaderni di Ricerca in Didattica (20), supplement n.1,* 411-424.

Oldenburg, R. (2012). Structure of Algebraic Proficiency. Research Paper ICME 12.

Sfard, A. (2000). Symbolizing mathematical reality into being: How mathematical discourse and mathematical objects create each other. In P. Cobb,K. E. Yackel, & K. McClain (Eds), *Symbolizing and communicating: perspectives on Mathematical Discourse, Tools, and Instructional Design* (pp. 37-98). Mahwah, NJ: Erlbaum.

Tourlakis, G. (2003) *Lectures in Logic and Set Theory Vol. 1 : Mathematical Logic*. Cambridge: Cambridge University Press.

Weber, K., & Alcock, L.: 2004, Semantic and syntactic proof productions, *Educational Studies in Mathematics, 56*, 209–234.