Language Performance at High School and Success in First Year Computer Science

Sarah Rauchas
School of Computer Science
University of the Witwatersrand, Johannesburg
Wits 2050, South Africa
sarah@cs.wits.ac.za

Benjamin Rosman
School of Computer Science
University of the Witwatersrand, Johannesburg
Wits 2050, South Africa
benjamin@cs.wits.ac.za

George Konidaris
Department of Computer Science
University of Massachusetts at Amherst
Massachusetts, USA
gdk@cs.umass.edu

Ian Sanders
School of Computer Science
University of the Witwatersrand, Johannesburg
Wits 2050, South Africa
ian@cs.wits.ac.za

ABSTRACT

We describe the first part of a study investigating the usefulness of high school language results as a predictor of success in first year computer science courses at a university where students have widely varying English language skills. Our results indicate that contrary to the generally accepted view that achievement in high school mathematics courses is the best individual predictor of success in undergraduate computer science, success in English at the first-language level in high school correlates better with actual performance. We discuss the implications of this for universities whose medium of teaching is English, operating in social contexts where many students are not native English speakers.

Categories and Subject Descriptors

K.3 [**Computers and Education**]; K.3.2 [**Computer and Information Science Education**]

General Terms

Performance

1. INTRODUCTION

The research presented here investigates the belief that language ability influences success in computer science at first year level. The most common criterion for success in computer science is believed to be aptitude for mathematics [5, 7, 11, 3]. Many universities throughout the world use high school mathematics results to select their students. It has also been suggested that success in a range of high school subjects is also a contributing factor to doing well in computer science at university [5]. Factors such as attitude to study, determination to succeed, self-efficacy, etc. are also important factors, but more difficult to accurately measure. While we have found that mathematics ability does play a role, we have come to believe that this is not the only factor, and might not even be the most important factor.

At our university, first year computer science consists of four topics: Basic Computer Organisation (BCO), Fundamental Algorithmic Concepts (FAC), Data and Data Structures (DDS) and Limits of Computing (LOC) [9]; all except one have almost no essay or paragraph writing component. Of these, we would expect that only the topic that does include essays (Limits of Computing, which discusses—as one of its components—social and ethical issues related to computing) would be perceived as challenging for students without a strong English ability. However, this has turned out not to be the case: we have been unable to significantly distinguish between this topic and the three more mathematical ones, which are concerned with algorithms, data structures, and computer organisation.

Recently we began to consider the idea that English ability is affecting students' capacity for succeeding in even the mathematically oriented topics. We looked at high school leaving examination results for English and mathematics and tried to see if there was any significant correlation between these and performance in first year computer science topics. Preliminary results indicated that there might be some correlation, so we decided to investigate this further. In particular we decided to broaden the language question to consider language more generally, rather than just English in high school final examinations.

This has involved two components: we surveyed our new intake of students, to gain some information about their comfort with language, their reading habits, and their perception of the importance of language in studying computer science; and we have done a more in-depth quantitative analysis of the high school examination results that we have access to, which includes looking at all language topics that students have taken. These can be seen, broadly, as

SIGCSE'06, March 1–5, 2006, Houston, Texas, USA.

in-course predictors and pre-course predictors respectively. With respect to the former, analysis of the more qualitative data, around language use and comfort, will be completed after a follow-up survey that looks at whether any of the habits or perceptions have changed while the students have tried to come to grips with their course material. In this paper we focus primarily on the latter—the performance in high school final examinations.

2. BACKGROUND

2.1 The South African schooling system

In South Africa, students do matriculation exemption examinations, referred to as matric exams, to obtain exemption from writing university entrance examinations. These are typically written at the end of 12 years of schooling, and in a variety of subjects. At least six subjects must be passed for the matric to be obtained. The subjects passed must include at least two of the 11 official languages.

All subjects can be taken at either Higher Grade (HG) or Standard Grade (SG) level. The difference between HG and SG is primarily in the depth and detail covered in the classes and examinations. Orthogonal to this, languages may be studied as first or second languages, each of which has a different focus. As a first language (FL) subject, the curriculum will include literature and poetry and be more analytical in its approach, while as a second language (SL) subject the focus will be more on grammar, vocabulary and the language itself. Native speakers of a language usually take it as an FL subject, and other languages as SL. Further, it is possible to study more than one language as FL, and more than one as SL. It is possible, for example, for a student to study Zulu FL, English FL, Afrikaans FL and Sotho SL. Any of these four might be at HG; equally any of these four might be at SG. It is most usual for students to take one language—the one they speak at home—as FL and another at SL; students who are hoping for a good matric will generally take both of these at HG.

2.2 Predicting success at first-year computer science

The problem of students not succeeding in computer science has been evident in many institutions, leading to a range of research trying to determine why students drop out of, or fail, computer science courses. Selection criteria for many institutions are based on high school mathematics results and an aggregate score based on the common view among educators that a student who does well in high school mathematics will also do well in computer science [5, 11, 7, 3]. At our university, we use high school mathematics as our primary criterion for admitting students into computer science courses. We have high drop-out and failure rates, which leads us to believe that the mathematics criterion is not serving our needs appropriately.

In contrast, a recent publication by Spark[11] claims a positive significant relationship between mathematics results at high school and performance in information systems. While this result was of interest to us, especially since it is based on students writing the same matric mathematics exams as ours, we believe that the difference between IS and CS at university level might play a role here. Also, Spark did not consider language in her study, and there is a possibility that if she did she might find it also a factor.

According to Campbell and McCabe[5], it is not appropriate to use a single high school subject as a predictor of success in computer science; a better indicator of success is an overall average of the high school results. This work was done some time ago, and seems now to more generally accepted; considering an aggregate is part of many institutions admission criteria. Older findings from Gathers[6] are also worth revisiting. In a wide-ranging study of around 250 students, ten factors were examined. Each factor that was suspected of having any influence on students' performance in computer science was investigated. The factors included high school grades and SAT score for each subject. The research showed that English is the best single predictor of success in computer science, with the overall best predictor of success in computer science being the combination of English and UTM mathematics placement score. To further establish that the results were valid a new cutting score equation was derived based on the research; this was used to admit students in the following year and it reduced computer science failure from 28% to 18%.

More recently, Barton and Neville-Barton[1] in Auckland, New Zealand, have investigated the connection between language and mathematics learning. Their work is also situated in a context where there are native English speakers as well as students whose first language is not English. Their results seem to indicate that there is a strong linguistic component to consider. Though their work says nothing explicit about computer science, it is worth considering whether there might be a connection, given that our computer science courses are strongly mathematically oriented. The other component, the mix of English proficiency, is also a strong property of the context in which we are based.

In South Africa, a large proportion of black students are not very fluent in English, as English is not their mother tongue. Research done by Nolan[8] at a South African institution, indicates that most students do not think they have any problem with English. However, even though students are confident about their English ability, there are indications that their actual ability may be less than they perceive it to be [4, as cited in [8]].

¿From the above, we wish to distill the factors that influence our research:

- There is a need to distinguish between performance in computer science, and more general computing related topics such as programming and information systems. Our focus here is on computer science, within the context of a mathematical orientation.

- The South African tertiary education system serves students from a wide range of backgrounds. Most of them are not native English speakers, yet at many universities, lecture and tutorial material is presented in English. There is also enormous language diversity, with the country having 11 official languages.

- While we certainly do not believe that performance at high school is the only factor influencing success at university, we believe it is part of the picture, and valuable as a cost-effective predictor.

3. DATA

Our entrance requirements for first year students are that they have obtained a C for mathematics at matric level at

Table 1: Correspondence between matric symbols and actual percentages.

Symbol	Percentage range	Midpoint used
A	80–100	90
B	70–79	75
C	60–69	65
D	50–59	55
E	40–49	45
F	0–39	20

HG. In South Africa, all pupils are required to pass English in order to matriculate, so all of our students will have either an SG or HG English pass, either as FL or SL.

We obtained data for our current class of first years, some of whom we also surveyed when they registered for computer science. For a sample of 107 students, we obtained the following data:

- matric results for: mathematics, English (FL), English (SL) and any other language (FL)
- university results for fundamental algorithmic concepts (FAC) and basic computer organisation (BCO)

Unfortunately the data we were able to obtain for the matric results was in classes (A, B, C, etc.) rather than actual results. The symbols relate to mark ranges as shown in Table 1. We've taken the mid-point of each range as the actual mark; this approach is not problem free, as discussed in Section 4.3. For the computer science topics we have actual results, with a normal distribution, in the range up to 100%.

Not all of our students are doing all of the four first year courses given. Some repeat courses they have failed, and others transfer courses from other Universities, though this is less common.

At this point in the year we only have results for the two courses mentioned. We will perform the same analysis on their results for the other courses that they will complete in November. For the two courses that have been completed, we performed various analyses on the data we obtained, and these are presented in the next section. For each analysis, we used only the collection of students for whom we had relevant data. For example, we wanted to analyse the connection between results for English FL and performance in BCO. To do this, we only used those students from the class who were doing BCO, and of those, only the ones who had done English FL at matric level. Thus the number of students in the sample varies between tests.

4. RESULTS

Our results are preliminary, and in Section 4.3 we discuss their limitations. However, there are some interesting outcomes, presented here.

4.1 Tests performed

We performed a Pearson product-moment correlation[10] on the data, for four categories: mathematics, English FL, English SL and all languages taken as FL. This was used to evaluate statistical significance.

For each of the above, we obtained Pearson's correlation figures (r) for the relation between FAC and BCO for the group of students who had taken that subject at high school. For each category, we also calculated r for the relation between FAC and the subject, and BCO and the subject. These are summarised in Table 2, together with the sample sizes (n). All figures with the exception of the two relating English SL to FAC and BCO respectively show statistical significance. There is a strong correlation between performance in BCO and FAC, in all cases. We found this reassuring, in that the courses—though taught by different lecturers—are believed to be of similar standard and difficulty.

We used only HG matric results. A well-accepted conversion between HG and SG results is to subtract 20% to 40% from an SG result to obtain its HG equivalent. We felt that this was too imprecise an option, and so discarded all results that were not at HG level. This also contributed to the varying sample sizes in the four categories listed above. For the language matric results, we kept FL and SL subjects separate, since the curricula each deals with are very different from each other. In the cases—there were only two of these—where more than one language was done as FL, we took the average for that student's matric results for those languages. Another important consideration is that our sample already has a preselection on the mathematics results; most of our students will have obtained at least a C for mathematics at HG; they only need to have obtained a pass (E or above) for their language and other subjects.

There is a weak positive correlation between matric mathematics results and performance in the computer science topics. Given that this is the primary predictor used thus far, a review of admission criteria is clearly due. Similarly, there is a weak positive correlation between matric FL results and performance in computer science. This suggests that general language competence is at least as good a predictor as mathematics.

More importantly, there is a much stronger statistically significant positive correlation between English FL results and computer science performance. On the other hand, the correlation between English SL and computer science performance was not statistically significant. This confirms our initial expectations, and indicates that further investigation along this thread is warranted. Although we have argued for many years that students whose home language is not English should still cope well within mathematically oriented subjects, where, for example, essay writing is not prioritised, this is not turning out to be the case.

Table 2: Pearson's correlation for the four categories.

Category	BCO to FAC	Category to FAC	Category to BCO	n
Mathematics	0.7607	0.2664	0.2782	90
English 1st language	0.7667	0.4571	0.4063	48
English 2nd language	0.7755	0.2343	0.1232	46
All 1st language	0.7651	0.3213	0.2440	94

4.2 FAC and BCO

An interesting observation is that though the mathematics high school results have similar Pearson's correlation figures for both BCO and FAC, for the other three categories—all of the ones that use language results from high school—there is a stronger correlation with the FAC performance than with the BCO performance. We plan to examine this aspect more closely, especially in the context of obtaining results for the other two computer science topics that the students complete in the latter part of the year. BCO and FAC are taught by different lecturers, but they also do have differing degrees of descriptive components, and we will look at whether this is a factor.

4.3 Further work

We believe that we could improve the analysis we have done on our data, to obtain a clearer picture. As mentioned before, obtaining actual results for the high school matric results, rather than using a midpoint, is important. We plan to do this as soon as possible.

Performing the same analyses for the other computer science courses, when the exams have been completed in November, will also give us additional information. In particular, we are interested in establishing whether there is any significant difference in the results for the Limits of Computing course, which has explicit essay components, compared with all the other courses which are believed to be more mathematically based.

We do note the differences in sample sizes between the four categories; however we believe that the trends are still relevant.

5. DISCUSSION

The results reported in this paper suggest that achievement in language courses at high school is a better predictor of success for our first year courses than mathematics at high school. More striking, and of more concern, is that English FL is a much better predictor. Like many other institutions, though we require competency in English as a prerequisite for entry into our courses, we have always claimed that we would not penalise those of our students who are not native English speakers. It seems as though this might not be true. At the very least, there appears to be an indication that students who do not do well in English FL at matric level are less likely to succeed in first year computer science. If this is true, we have an obligation to identify students at risk and establish ways in which to help them.

It is however very important to point out that we believe there is a bigger picture, and we wish to investigate this further. The bigger picture that we're interested in is the notion that it's not just about the particular language—English in this case—but about appreciation of language and its use.

Anecdotally, we have found that there seems to be a distinction between students who appreciate precise language use, and those who don't. For example, given the sentence *"What properties do lists have that dictionaries don't?"* in the context of a discussion on data structures, we noticed a range of responses from students. Some students would give properties that lists have, but which dictionaries also have. Some students would give things that were not actually properties, such as ways to access elements in lists. When talking to the students about their answers, many of them could not see why those answers were not well related to the question.

It is a widely held view that because our students are learning technical concepts, there is no advantage to their being more comfortable with English. Mathematical and theoretical concepts are taught in lectures, and the language used is introduced in this scenario. It's new to both native and non-native speakers of the delivery language. However, there are two ways in which experience with the language of instruction is an advantage. First, it's easier for students who are already comfortable with the language to learn any new words. It's quicker to look things up in the dictionary, and there is the chance that they may already be familiar with the words. Secondly, it's easier to make sensible guesses about what a new word might mean, and it's also easier to remember what words signify, through their connection with other known words.

An anecdote illustrating this is the question where students were asked to discuss something about the *initial part* of a list. A number of students simply did not know what *initial* meant. We argue that someone who is already comfortable with English, but doesn't know what initial means in the context of a data structure, could use their knowledge of the language to work it out. For example, they might think that their own initials are the first letters of each of their names, and would be able to transfer this to the context of a list. Related to this is the notion that language is something that one can play about with, and figure out, rather than being some magical thing that an oracle knows. We argue further that this approach to taking control of language is not specific to being a native English speaker, but rather to do with the relationship one has with language in general. This points to the need to look at more general aspects of language use, such as comfort with language, interest in reading more generally (novels and non-fiction, as well as academic writing) and television watching habits (and which language channels students tend to choose to watch). We have included such aspects in our survey mentioned earlier.

6. CONCLUDING REMARKS

Although it would be simplistic to expect that language ability is the only factor affecting students' performance in first year computer science, we are finding that some aspects of language ability might be a better predictor of success in first year than high school mathematics results. Nevertheless, there are many other factors that we still need to take account of.

We believe that the picture is not simply about how students have done in their high school exams in language, and especially in English. What we are more interested in is their comfort with language in general, and their appreciation that language use is a precise art—this isn't measured by their results at high school. Furthermore, we believe that there is a component to do with transition that might also play a role. As well as coming to grips with the actual material of computer science at university, there is a very different approach to learning that students need to take on board during their first year away from high school. Part of this is an appreciation of the role of language in communicating ideas and concepts, rather than fluency in a particular language.

Finally, we need to say something about the content and

orientation of our computer science courses. There is a view that the way that teaching happens in the sciences is very different from how it happens in the humanities. In particular, Buckland[2] argues that in the sciences, the focus is much more on factual knowledge and quantitative approaches, while in the humanities, deep understanding is emphasised. In our department, however, we do favour deep understanding over factual content and technical ability. This is something that many of our students struggle with, in the transition from high school to university. One thought from this is that our courses might be different from the other science courses that our students are doing at first year level, and ours might be more in line with the approaches Buckland suggests are taken in the humanities. This could indicate that those students who are comfortable with language, and other aspects of humanities study, cope better with the deep understanding emphasis we put into our computer science courses.

In summary, the results we have obtained from analysing matriculation results and the relation to performance in our first year course do indicate that there is a language component worth investigating further. It is a complex one in the context of a group of students with very different language backgrounds, and also very different academic abilities. Further investigation should be both quantitative and also qualitative, and is vital to understanding how to help our students to succeed, and also in identifying those who are more likely to do so.

7. ACKNOWLEDGMENTS

Thanks to Alexander Holt for assistance with proofreading, sanity checking and typesetting, and to Vashti Galpin for discussions about our statistical analysis.

This work was partially supported through a Teaching and Learning Research Grant from the University of the Witwatersrand, for which we are very grateful.

8. REFERENCES

[1] Bill Barton and Pip Neville-Barton. Investigating the relationship between English language and mathematical learning. In *CERME 3, Third Conference of the European Society for Research in Mathematics Education*, Italy, 2003.

[2] Richard Buckland. Can we improve teaching in Computer Science by looking at how English is taught? In *Proceedings of the Second Australasian Conference on Computer Science Education*, pages 155–162. ACM Press, 1996.

[3] D. F. Butcher and W. A. Buth. Predicting performance in an introductory Computer Science course. *Communications of ACM*, 28:263–268, 1984.

[4] Q. Buthelezi. Black South African English. *Language and Social History: Studies in South African Sociolinguistics*, pages 242–250, 1995.

[5] Patricia F. Campbell and George P. McCabe. Predicting the success of freshmen in a computer science major. *Communications of ACM*, 27(11):1108–1113, 1984.

[6] E. Gathers. Screening freshmen Computer Science majors. *ACM SIGCSE Bulletin*, 18(3):44–48, September 1986.

[7] Annagret Goold and Russel Rimmer. Factors affecting performance in first-year computing. *ACM SIGCSE Bulletin Inroads*, 32(2):39–43, June 2000.

[8] Verena Nolan. Influence of attitude towards statistics, English language ability and mathematical ability in the subject quantitative techniques at the Vaal Triangle Technicon, South Africa, 2002. Vaal Triangle Technicon, South Africa.

[9] I. D. Sanders and C. S. M. Mueller. A fundamentals-based first year computer science curriculum. In *Proceedings of the 31st Special Interest Group on Computer Science Education Technical Symposium*, pages 227–231, Austin, Texas, 2000. SIGCSE.

[10] David J. Sheskin. *Handbook of parametric and nonparametric statistical procedures.* Chapman and Hall/CRC, 2000.

[11] L. Spark. Matric maths as a predictor of success in Information Systems — a study of Information Systems students at the University of the Witwatersrand. In *Proceedings of the 35th Conference of SACLA*, pages 268–271, Kasane, Botswana, 2005.